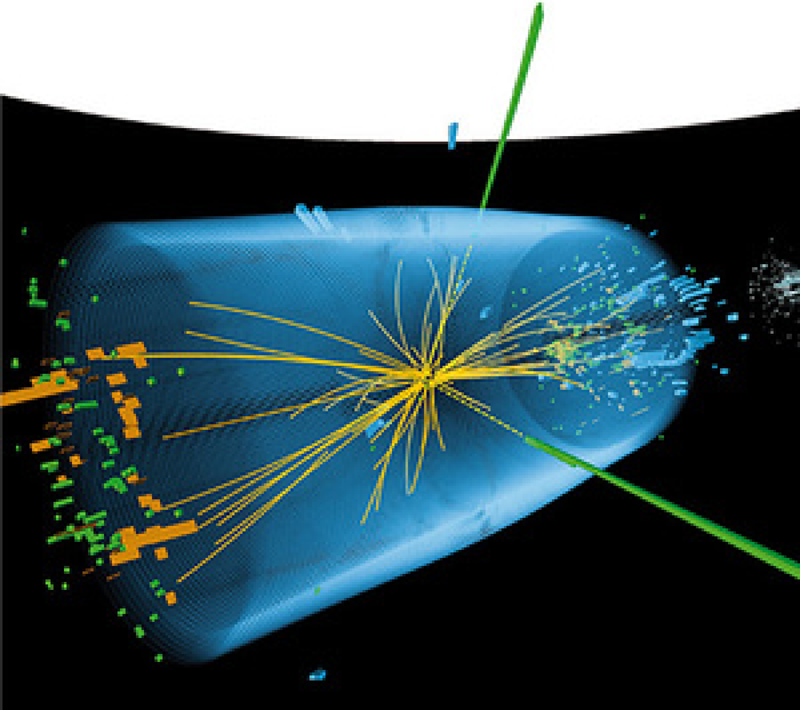
Yorikiyo Nagashima

Beyond the Standard Model of Elementary Particle Physics



Yorikiyo Nagashima

Beyond the Standard Model of Elementary Particle Physics

Related Titles

Narsky, I., Porter, F.C.

Statistical Analysis Techniques in Particle Physics

Fits, Density Estimation and Supervised Learning

2014
ISBN: 978-3-527-41086-6
Also available in digital formats.

Behnke, O., Kröninger, K., Schott, G., Schorner-Sadenius, T. (eds.)

Data Analysis in High Energy Physics

A Practical Guide to Statistical Methods

2013
ISBN: 978-3-527-41058-3
Also available in digital formats.

Brock, I., Schörner-Sadenius, T. (eds.)

Physics at the Terascale

2011
ISBN: 978-3-527-41001-9
Also available in digital formats.

Hauptman, J.

Particle Physics Experiments at High Energy Colliders

2011
ISBN: 978-3-527-40825-2

Mandl, F., Shaw, G.

Quantum Field Theory 2e

2nd Edition

2010
ISBN: 978-0-471-49684-7
Also available in digital formats.

Stock, R. (ed.)

Encyclopedia of Applied High Energy and Particle Physics

2009
ISBN: 978-3-527-40691-3

Röser, S. (ed.)

Reviews in Modern Astronomy

Vol. 18: From Cosmological Structures to the Milky Way

2005
Print ISBN: 978-3-527-40608-1
Also available in digital formats.

Lilley, J.

Nuclear Physics – Principles & Applications

2001
ISBN: 978-0-471-97935-7
Also available in digital formats.

Thomas, A.W., Weise, W.

The Structure of the Nucleon

2001
ISBN: 978-3-527-40297-7
Also available in digital formats.

Yorikiyo Nagashima

Beyond the Standard Model of Elementary Particle Physics

WILEY-VCH
Verlag GmbH & Co. KGaA

The Author

Yorikiyo Nagashima
Osaka University
Japan
nagashimayori_1128@yahoo.co.jp

Cover
© 2012 CERN, for the benefit of the CMS Collaboration

All books published by **Wiley-VCH** are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

© 2014 Wiley-VCH Verlag GmbH & Co. KGaA, Boschstr. 12, 69469 Weinheim, Germany

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Print ISBN: 978-3-527-41177-1
ePDF ISBN: 978-3-527-66505-1
ePub ISBN: 978-3-527-66504-4
Mobi ISBN: 978-3-527-66503-7
oBook ISBN: 978-3-527-66502-0

Cover Design Adam-Design, Weinheim, Germany
Typesetting Laserwords Private Limited, Chennai, India
Printing and Binding Markono Print Media Pte Ltd., Singapore

Printed on acid-free paper

Contents

Preface *XIII*

Acknowledgments *XVII*

Glossary *XIX*

1 Higgs *1*
1.1 Introduction *1*
1.2 Higgs Interactions *2*
1.2.1 Standard Model *2*
1.2.2 Lagrangian After Symmetry Breaking *5*
1.2.3 Decay Modes *6*
1.3 Mass *9*
1.3.1 Predictions from EW Data *9*
1.3.2 Vacuum stability *10*
1.3.3 Theoretical Upper Limit *12*
1.4 Little and Big Hierarchy Problem *16*
1.5 Higgs in the Supersymmetry *19*
1.5.1 Two Higgs Doublets *19*
1.5.2 Coupling Strengths of MSSM Higgs *22*
1.5.3 Mass Spectrum of MSSM Higgs *24*
1.6 Is the Higgs Elementary? *28*
1.6.1 Technicolor Model *29*
1.6.2 Little Higgs Model *31*
1.7 Production and Detection of Higgs *38*
1.7.1 Higgsstrahlung $e^-e^+ \to hZ$ *39*
1.7.2 W Boson Fusion *39*
1.7.3 Productions at the Hadron Collider *43*
1.7.4 Signals at LHC *46*
1.7.5 Higgs Detection Methods *48*
1.7.6 Discovery of Higgs *51*
1.7.7 SM Higgs? *52*

1.7.8 MSSM Higgs and Future Prospect *55*
1.8 Summary *58*

2 Neutrino *61*
2.1 Introduction *61*
2.2 Neutrino Mass *62*
2.2.1 Mass Matrix *65*
2.2.2 Left-Right Symmetric Model *69*
2.3 Electromagnetic Interaction *70*
2.4 Neutrino Mixing *73*
2.5 Neutrino Oscillation *76*
2.5.1 Two-Flavor Oscillation *76*
2.5.2 Atmospheric Neutrino *79*
2.5.3 Accelerator Experiments *82*
2.6 Underground Detectors *87*
2.7 Solar Neutrino *93*
2.7.1 The Solar Puzzle *93*
2.7.2 Matter Oscillation *100*
2.7.3 Reactor Experiment *111*
2.8 Three-Flavor Oscillation *114*
2.8.1 PMNS Matrix *114*
2.8.2 Summary of Experimental Data *120*
2.8.3 CP Violation and Mass Hierarchy *121*
2.8.4 Future Prospects *124*
2.9 Double Beta Decay *126*
2.9.1 The Effective Majorana Mass *128*
2.9.2 Current Status *129*
2.9.3 To Design an Experiment *131*
2.9.4 Experimental Apparatus *133*
2.10 Supernova Neutrino *136*
2.10.1 Stellar Evolution *138*
2.10.2 Feedback to Particle Physics *150*

3 Grand Unified Theories *155*
3.1 Introduction *155*
3.2 Why GUTs? *155*
3.2.1 Weinberg Angle in GUTs *157*
3.2.2 Quantization of the Electric Charge *157*
3.2.3 Triangle Anomaly *158*
3.3 $SU(5)$ *160*
3.3.1 Fermion Representation *161*
3.3.2 Representation of the Gauge Particle *164*
3.3.3 Symmetry Breakdown *168*
3.3.4 Predictions *170*
3.4 $SO(10)$ *174*

3.4.1 Left–Right Symmetric World *174*
3.4.2 New Gauge Bosons Z' and W' *175*
3.5 Hierarchy Problem *182*
3.6 SUSY GUT *185*

4 Supersymmetry I: Basics *189*
4.1 Introduction *189*
4.1.1 Toy Model *190*
4.1.2 Field Theoretical Operators *191*
4.2 Two-Component Formalism *193*
4.2.1 Majorana Fields *193*
4.2.2 SUSY Operators *198*
4.2.3 Superspace *200*
4.3 Chiral Superfield *203*
4.3.1 Products of Chiral Superfields *206*
4.4 Vector Superfields *206*
4.4.1 Field Strength *209*
4.5 Action *210*
4.5.1 SUSY Invariant Action *210*
4.5.2 Kinetic Energy of Chiral Superfield *212*
4.5.3 Superpotential *213*
4.5.4 Lagrangian of the Chiral Fields *215*
4.5.5 Kinetic Energy of Vector Field *216*
4.6 Gauge Interaction *217*
4.6.1 Global $U(1)$ Transformation *217*
4.6.2 Local $U(1)$ Transformation *217*
4.6.3 Non-Abelian Interaction *219*
4.7 Summary of SUSY Lagrangian *220*
4.8 Spontaneous Symmetry Breaking *221*
4.8.1 D-Term Breaking *222*
4.8.2 F-Term Breaking *223*

5 Supersymmetry II: Phenomenology *225*
5.1 Introduction *225*
5.2 Minimum Supersymmetric Standard Model *226*
5.2.1 Particle Spectrum *226*
5.2.2 Interactions *229*
5.2.3 Constraints *230*
5.2.4 SUSY Breaking *231*
5.2.5 Higgs Potential *232*
5.3 Minimum SUGRA *235*
5.3.1 Soft- SUSY Breaking *235*
5.3.2 Mass Formula *237*
5.3.3 μ Problem *241*
5.4 GMSB *241*

5.4.1 Messenger Particles *242*
5.4.2 Mass Formula *242*
5.4.3 Features of GMSB *244*
5.5 AMSB and Extra Dimension *245*
5.6 Summary of Mass Spectra *247*
5.7 Searches for Sparticles *248*
5.7.1 Production Mechanism *249*
5.7.2 Sleptons *250*
5.7.3 Charginos and Neutralinos *253*
5.7.4 LSP *255*
5.7.5 Gluino and Squarks *255*
5.7.6 Stop *260*
5.7.7 R-hadrons *261*
5.7.8 Gravitino *262*
5.8 Current Status *263*

6 Extra Dimension *267*
6.1 Introduction *267*
6.2 KK Tower *270*
6.2.1 Effective Coupling Strength in 4D *272*
6.3 Chiral Fermions *273*
6.3.1 Orbifold S^1/Z_2 *275*
6.3.2 Mass Generation and Localization *278*
6.3.3 Hierarchy *282*
6.3.4 Split Fermion Scenario *285*
6.4 Gauge Field in ED *287*
6.4.1 Action in 4D *287*
6.4.2 Coupling Strength *288*
6.4.3 Gauge–Higgs Unification *289*
6.5 Gravitational Field *292*
6.5.1 Decomposition of the Gravitational Fields *294*
6.6 Warped Extra Dimension *296*
6.6.1 Anti-de Sitter Space AdS_5 *296*
6.6.2 RS1 Scenario *299*
6.6.3 RS2 scenario *300*
6.6.4 Gravitons in the RS Model *302*
6.6.5 Signals for Warped ED *305*
6.7 Universal Extra Dimension (UED) *308*
6.7.1 General Features *308*
6.7.2 Selection Rules *308*
6.7.3 Constraints *311*
6.7.4 Signals for UED *312*
6.8 Searches for Generic ED *313*
6.8.1 Astrophysical Constraints on ADD Models *314*
6.8.2 Collider Experiments on ADD Models *316*

6.8.3 TeV^{-1} Extra Dimension Model: *322*
6.9 Black hole production *325*

7 Axion *329*
7.1 Soliton *329*
7.1.1 Kink *329*
7.1.2 Vortex *334*
7.1.3 Winding Number *337*
7.1.4 Spacetime Where the Soliton Lives *339*
7.1.5 Instanton *340*
7.1.6 θ Vacuum *347*
7.1.7 Electroweak Vacua *348*
7.2 Strong CP Problem *350*
7.2.1 Anomaly *350*
7.2.2 Chiral Transformation and the Mass Term *353*
7.2.3 U(1) problem *355*
7.3 Why Do We Need the Axion? *356*
7.3.1 PQ Symmetry and the Standard Axion *357*
7.3.2 Invisible Axion *360*
7.4 Constraints on Invisible Axions *363*
7.4.1 Coolant of the Stellar Evolution *363*
7.4.2 Axion as the Dark Matter *367*
7.4.3 Misalignment axion *367*
7.5 Laboratory Axion Searches *370*

8 Cosmology I: Big Bang Universe *377*
8.1 Why Do We Study Cosmology? *377*
8.2 Cosmic Equation *378*
8.2.1 Robertson–Walker Metric *379*
8.2.2 Friedmann Equation *382*
8.3 Expanding Universe *385*
8.3.1 Redshift of Light *385*
8.3.2 Redshift of Particles *386*
8.3.3 Cosmic Parameters *386*
8.4 Thermal Universe *389*
8.4.1 Thermodynamics *389*
8.4.2 Radiation and Matter Dominance *392*
8.4.3 Time versus Temperature *393*
8.4.4 Overview of Thermal History *394*
8.5 Cosmic Distance, Horizon *396*
8.5.1 Distance *396*
8.5.2 Horizon *399*
8.6 Genesis *400*
8.6.1 Matter Universe *400*
8.6.2 Baryogenesis *401*

8.6.3 Leptogenesis *404*
8.6.4 Neutrino Decoupling *408*
8.6.5 Big Bang Nucleosynthesis *410*
8.7 Last Scattering *414*
8.7.1 Radiation–Matter Equality *414*
8.7.2 Recombination *415*
8.7.3 Dark Age *417*
8.8 Inflation *418*
8.8.1 Slow Rolling and Reheating *418*
8.8.2 Horizon Problem *421*
8.8.3 Flatness Problem *423*
8.8.4 Monopole Problem *424*

9 Cosmology II: Structure Formation *425*
9.1 Galaxy Distribution *425*
9.1.1 Introduction *425*
9.1.2 Boltzmann Equation *429*
9.1.3 Growth of the Fluctuation *434*
9.1.4 Dark Matter *435*
9.1.5 Jeans Wavelength of the Neutrino *436*
9.1.6 Power Spectrum *437*
9.1.7 Initial fluctuation *444*
9.1.8 Effects of Neutrino Mass *448*
9.1.9 Primordial Fluctuation *448*
9.2 CMB Anisotropy *454*
9.2.1 Overview *454*
9.2.2 Sachs–Wolfe Effect *458*
9.2.3 Acoustic Oscillations *459*
9.2.4 Doppler Effect *462*
9.2.5 Silk Damping *462*
9.2.6 Outcome of CMB Measurements *464*
9.2.7 Polarization *467*

10 Dark Matter *475*
10.1 Cosmic Budget *475*
10.2 Evidences of Dark Matter *475*
10.2.1 Rotation Curves of Spiral Galaxies *476*
10.2.2 Virial Mass of the Clusters *477*
10.2.3 X-ray Emitting Clusters *478*
10.2.4 Gravitational Lens *479*
10.3 Relics of the Big Bang *489*
10.3.1 Freeze-Out *489*
10.3.2 Hot Dark Matter *490*
10.3.3 Cold Dark Matter *491*
10.3.4 Candidates for the Dark Matter *493*

10.4 How to Detect? 495
10.4.1 Indirect methods 495
10.4.2 Production by Accelerators 497
10.4.3 WIMPS Wind 498
10.5 Searches for DMs in the Halo 505
10.5.1 General 505
10.5.2 Bolometer 507
10.5.3 Xe Detector 508
10.5.4 Current Status 512

11 Dark Energy 513
11.1 Dark Energy 513
11.1.1 Accelerating Universe 513
11.1.2 Cosmic Age 515
11.1.3 ΛCDM Model 519
11.2 Cosmological Constant 520
11.3 Quintessence model 523
11.4 Other Dark Energy Models 532
11.5 How to Investigate the Dark Energy? 533

Appendix A Virial Theorem 543

Appendix B Chandrasekhar Mass 545

Appendix C Production of KK Gravitons 549

Appendix D Homotopy 551

Appendix E General Relativity 559
E.1 Geodesic Equation 559
E.2 Ricci Tensor and Scalar 561
E.3 Gauge Degrees of Freedom 561
E.4 Gravitational Waves 563

Appendix F Tensor Spherical Harmonic Function 565

Appendix G Destiny of the Cosmos 567

Appendix H Answers to Some Problems 571

References 575

Color Plates 597

Index 617

Preface

Modern particle physics started in 1935 when Fermi and Yukawa proposed theories of weak and strong interactions, respectively. The 40-year saga in the quest for the ultimate form of matter and the interactions that govern them culminated in the Standard Model (SM) of particle physics in the early 1970s.

Nearly 50 years have passed since the SM was established. It is a miracle that it still holds the status as the ultimate theory of matter at the most fundamental level. No experimental observations that contradict the SM have been discovered, with perhaps one exception. Even the neutrino oscillation, the exception, may be considered as a small extension of the SM that does not need modifications. As the theory of relativity and quantum mechanics were born as a result of searches beyond Newtonian mechanics and electromagnetism, we expect that a new physics exists beyond the SM.

The SM established a prescription to unify forces by way of the gauge symmetry and spontaneous symmetry breaking. The grand unified theories (GUTs), the super-gravity (SUGRA), and the string theories were developed as extensions of the SM. Problems were pointed out and hints and new ideas have been suggested in developing the unified theories. The hierarchy is an outstanding problem among them. Many theoretical ideas including the super-symmetry (SUSY) and the extra dimension (ED) have been proposed to solve the problem. Most of them suggest a new physics at the teraelectronvolt (TeV) energy scale.

It has also been pointed out from the very beginning that the SM will lose its predictive power on phenomena beyond the TeV energy range (or $\sim 10^{-19}$ m in size). This is because the dynamics of the Higgs that causes electroweak phase transition below $\sim$ 1 TeV is unknown. The SM also established a notion that discovery of a new particle is synonymous with the discovery of a new physics. Therefore, experimental searches for new particles in the hitherto unexplored energy region, especially in the TeV range, are the most orthodox way to explore the physics beyond the SM.

On the other hand, experiments at energy scale in the range $\sim 10^{10-19}$ GeV are required to probe the physics of the unified theories. They are beyond the reach of present-day technology. Fortunately, the advent of the unified theories found a way to elucidate the history of early universe. Cosmology and particle physics have become one and the same scientific field. Conversely, the advent of cosmology opened a new window to view and probe the high-energy phenomena that are inaccessible by today's technology. We can probe properties of particles by looking at cosmic relic particles, fossils of the Big Bang, so to speak. Nowadays, researches in particle physics that do not rely on accelerators occupy an important branch. They are generically referred to as *non-accelerator physics*. It is an unappealing name, nonetheless used for the reason that no other has been invented.

This book is the third of a series of textbooks on ''Elementary Particle Physics'' [1] and [2]. Part 1 of Volume I [1] introduced the quantum field theory at the level that is necessary to understand phenomenology and to derive at least tree-level formulas for various reactions. Part 2 of Volume I described a way, logically as well as historically, to reach the SM of particle physics. Content of Volume II [2] is the SM itself, an essential part of the electroweak interactions, and quantum chromodynamics (QCD). This book, entitled ''Beyond the Standard Model of Elementary Particle Physics'' should be considered as Volume III. The title speaks of its content by itself but the discussions are limited to topics that will become experimentally accessible in the near future. Each of the three volumes is organized to stand on their own depending on the readers' interest and level, except occasional references to equations that were derived previously. This book is organized as follows:

Chapter 1 describes the properties of the Higgs particle, which is within the SM but whose dynamics is unknown. The Higgs mechanism constitutes the basis of the SM. Clarification of its dynamics may consolidate or modify the foundation of the SM. Therefore, it is the most urgent topic. The large hadron collider (LHC) was built primarily to discover the Higgs and clarify the underlying mechanism. It is now producing data.

Chapter 2 discusses the neutrino. Vanishing of the neutrino mass has always been the most outstanding topic from the beginning of its prediction by Pauli. The neutrino oscillation is a firm evidence that the neutrino is not massless, which is a topic on its own, and at the same time has provided the first and so far the only experimental data that goes beyond the SM. Clarifying relations among the three types of the neutrinos is the most direct approach to the flavor mystery, which is one of the unsolved problems of the SM. Furthermore, the neutrino provides important links to connect the SM with the GUTs and also with cosmology.

Chapter 3 discusses a basic structure of the GUTs. It does not aim to go deep in its structure but is meant to introduce problems that one encounters in trying to unify the forces. Above all, it has placed the so-called gauge hierarchy in limelight. The GUT is a topic by itself, but this chapter also serves as a prelude to the SUSY and the extra-dimensional theories.

Chapter 4 introduces basic algebra of SUSY that is necessary to derive formulas for the SUSY phenomenology. If the reader is interested only in the SUSY phenomenology, he or she may skip this chapter except the first introductory remarks and go directly to Chapter 5.

Chapter 5 discusses the phenomenology of the low-scale SUSY and how they are being examined by LHC data. The SUSY was offered as a remedy to solve the technical difficulties associated with the hierarchy problem. It also has a virtue that the gauged SUSY, referred to as SUGRA, can handle gravity and hence is a candidate for the unification of all forces. The super-string theory, the prime candidate for the ultimate unified theory, considers strings as fundamental building blocks of matter. It also respects the SUSY and works in 10- or 11-dimensional spacetime, which inspired the theory of the ED. However, the low-scale SUSY is treated as an independent phenomenological theory to solve the difficulties associated with the SM or its extension. It also offers a prime candidate for the dark matter (DM) in the universe.

The ED discussed in Chapter 6 provides an alternative to the SUSY to solve the hierarchy problem. Unlike the SUSY, which only solves the technical aspect of the hierarchy problem, ED offers a possible solution to deny the existence of the hierarchy itself, that is, that the energy scale of gravity may not be $1/\sqrt{G_N} \sim 10^{19}$ GeV but in the TeV region. Another interesting possibility is the gauge–Higgs unification, which might provide a symmetry to circumvent the hierarchy problem.

Chapter 7 discusses the axion and explains instantons, chiral anomaly, and the strong charge parity (CP) problem, which are in the realm of the SM but offer completely different aspects not provided by the perturbation theories. Existence of the axion seems an unavoidable outcome of these theoretical issues. It also provides an alternative candidate for the DM and a link between particle physics and cosmology. The first half of the chapter is devoted to these theoretical problems. Readers who are interested only in the phenomenology of the axion may skip this part.

Chapters 8 and 9 are devoted to cosmology. Chapter 8 describes a thermal history of the Hot Big Bang universe as a uniform and isotropic perfect fluid. Connections between particle physics and cosmology/astro-particle physics, including the inflation, are discussed. Chapter 9 deals with the deviation of matter distribution from uniformity, namely the large-scale structure of the universe, CMB (cosmic microwave background), and roles of the DM in forming them. Although these topics deserve treatments of their own, they are, in the author's mind, necessary introductions to tackle the problem of the DM and the dark energy (DE).

Chapter 10 discusses evidences and searches for the DM and possible candidates from the particle physics point of view. Finally, Chapter 11 discusses the DE. The field of DE is in its infancy. However, the author believes that both the DM and the DE will be the two main themes of particle physics in the twenty-first century.

Target readers of this book are experimental physicists, graduate students aiming at theories or experiments, and hopefully laymen in the field who are serious enough to follow the mathematical logics described in the book. They are expected to have a basic knowledge of particle physics at the level described in [1]. But they may be ignorant of each topic adopted in this book. Each chapter is basically independent and stands on its own, except Chapter 9 (Cosmology II) which quotes many results from the previous chapter. Some readers may also find it easier to read some parts of Chapter 8 (Cosmology I) first for understanding the invisible axions described in Chapter 7.

LHC made a historic discovery of the Higgs particle with mass 125 GeV in 2012. No indications of new physics have been found so far (as of summer 2013). The SM turned out to be much better than expected by many theorists. The validity of the

SM is now extended at least by a factor 10 in the energy scale to $\sim O(10\ \text{TeV})$. People began to cast doubt about the naturalness, which has been the guiding principle in proposing new models. This book faces the danger of becoming obsolete soon if LHC makes another revolutionary discovery during the rest of its operations. The most likely chapters to be affected by it are Chapters 1 and 5, which describe the Higgs and/or SUSY. Otherwise, the author hopes that most contents in this book will keep their usefulness longer. The author's optimism and decision to publish this book at this time are based on the following notions.

Usually, reviews of forefront topics are best provided by specialists in the field because their contents change fast and keeping track of the most recent idea/data without making mistakes is hard for nonspecialists. Indeed, if you search for available books by the title 'Beyond the Standard Model', you will find most of them are conference reports. The reason the author dared to challenge these topics is as follows. For some class of readers, organized reviews of frontier fields in a consistent style by the same author may have some merit of its own. As time goes, the data may become somewhat obsolete, but the basic concepts to pursue the subject will hopefully last longer unless its central idea is drastically changed by revolutionary discoveries.

Acknowledgments

The author would like to express his gratitude to the authors cited in the text and to the following publishers for permission to reproduce various figures and tables:

American Astronomical Society, publisher of the Astrophysical Journal for permission to reproduce Figures 2.24, 2.56b, 6.14, 9.1, 9.2, 9.6a, 10.1, 10.5, 11.2, 11.3, 11.4b, 11.6a, 11.14a and Table 10.1.

American Physical Society, publisher of the Physical Review, Physical Review Letters and the Review of Modern Physics, for permission to reproduce Figures 1.12a,b, 2.9, 2.12, 2.13a, 2.18, 2.19, 2.21a,b, 2.26a,b, 2.29a,b, 2.30, 2.32b, 2.33a,b, 2.37, 2.39a,b, 2.40a,b, 2.48a,b, 2.49a,b, 3.7a,b, 5.14a,b, 5.15a,b, 6.11, 6.12, 6.21a,b, 7.12b, 10.11, 10.12a,b, 10.20b, 10.21a, 10.23b, 11.13b, and 11.15.

Annual Reviews, publisher of Annual Review of Nuclear and Particle Science and Annual Review of Astronomy and Astrophysics for permission to reproduce Figures 2.25b, 2.54, 9.14, 10.8, 11.1a,b, and 11.12b.

Elsevier Science Ltd., publisher of Astroparticle Physics, Nuclear instruments and Methods, Nuclear Physics, Physics Letters, Physics Report, Progress of Particle and Nuclear Physics for permission to reproduce Figures 1.4, 1.6b, 1.9, 1.10a, 1.15b, 1.18a,b, 1.28b, 2.1a, 2.6, 2.16a, 2.20, 2.31b, 2.58, 3.11, 6.15a,b, 7.9, 9.7a,b, 10.10, 10.16a, 10.17b, 10.22b, 10.23a, 10.24, and 11.11a,b.

Institute of Physics Publishing Ltd., publisher of Journal of Instrumentation, Physica Scripta, and Report on Progress in Physics for permission to reproduce Figures 2.31a, 2.45, 2.47a,b, and 8.6.

Particle Data Group, publisher of Review of Particle Physics for permission to reproduce Figures 1.2a,b, 1.6a, 1.17, 1.27a, 2.8b, 2.13b, 2.15, 2.27, 5.8a,b, 6.1, 7.10, 7.11, 8.7, 9.6b, and 9.8.

Springer, publisher of European Journal of Physics and Journal of High Energy Physics for permission to reproduce Figures 1.11, 1.29, 5.13b, 6.18a,b, 6.23a,b, and 9.10a.

World Scientific, publisher of International Journal of Modern Physics for permission to reproduce Figures 2.42, and 10.18a,b.

Glossary

ACT	Atacama Cosmology Telescope
AD	anti-neutrino detector
ADD	Arkani-Hamed–Dimopoulos–Dvali
ADMX	Axion Dark Matter eXperiment
AFTA	Astrophysics Focused Telescope Assets
AGB	asymptotic giant branch
ALPs	axion-like particles
AMSB	anomaly mediated symmetry breaking
ATLAS	A Toroidal LHC ApparatuS
BAO	baryon acoustic oscillation
BB	beta-beam
BBN	Big Bang nucleosynthesis
BE	Bose–Einstein
BH	black hole
BOSS	Baryon Oscillation Spectroscopic Survey
Bq	Becquerel
BR	branching ratio
BS	blue stragglers
CAST	CERN Axion Solar Telescope
CC	charged current
CCD	charge-coupled device
CDM	cold dark matter
CDMS	cryogenic dark matter search
CGH	central galactic halo
CKM	Cabibbo–Kobayashi–Maskawa
CL	cluster distributions
CMB	cosmic microwave background
CMBPOL	CMB polarization
CMS	compact muon solenoi
CMSSM	constrained minimum supersymmetric extension of the Standard Model
CNGS	CERN neutrino to Gran Sasso

CNMSSM	constrained next MSSM
CNO	carbon-nitrogen-oxygen
COBE	cosmic microwave background explorer
CP	charge parity
CPT	conjugation–parity transformation–time reversal
DE	dark energy
DFSZ	Dine-Fischler-Srednicki-Zhitnitsky
DGLAP	Dokshitzer–Gribov–Lipatov–Altarelli–Parisi
DM	dark matter
ECC	*emulsion cloud chamber*
ED	extra dimension
ER	extremely relativistic
ES	elastic scattering
ESA	European Space Agency
ETC	extended technicolor
EW	electroweak
EWPD	electroweak precision data
EXO	enriched Xenon observatory
FC	fully contained
FCNC	flavor-changing neutral current
FD	Fermi–Dirac
FOM	Figure of Merit
GALLEX/GNO	Gallium Experiment/Gallium Neutrino Observatory
GC	globular cluster
Gd-LS	gadolinium-doped liquid scintillator
GeV	giga-electonvolt
GGM	general gauge mediation
GIM	Glashow–Iliopoulos–Maiani
GL	gravitational lensing
GMSB	gauge mediated symmetry breaking
GPS	global positioning system
GR	general relativity
GRB	gamma ray burst
GUT	grand unified theory
GW	giga-watt
HB	horizontal branch
HDM	hot dark-matter
H.E.S.S.	High Energy Stereoscopic System
HFETs	heterostructure field-effect transistors
HST	Hubble space telescope
IAV	inner acrylic vessel
IB	inner balloon
ID	inner detector

IGM	intergalactic medium
IMB	Irvine Michigan Brookhaven
IH	*inverted hierarchy*
ILC	intrernational electron collider
IR	infrared
IS	inverted spectrum
ISM	*interstellar medium*
ISW	*integrated Sachs–Wolfe*
IWS	inner water shield
JAXA	Japan Aerospace Exploration Agency
JDEM	joint dark energy mission
K2K	KEK to Kamioka
KamLAND	Kamioka Liquid Scintillator Anti-Neutrino Dertector
KamLAND-Zen	KamLAND zero neutrino experiment
keV	kilo-electronvolt
KK	Kaluza–Klein
KSVZ	Kim-Shifman-Vainstein-Zakharov
LAAPD	large-area avalanche photodiodes
LAMPF	Los Alamos Meson Physics Facility
LED	large extra dimension
LEP	large electron–positron
LH	left-handed
LHC	large hadron collider
LKP	lightest KK particle
LMA	large mixing angle
LQCD	lattice quantum chromodynamics
LS	liquid scintillator
LSST	Large Synoptic Survey Telescope
LSTC	low-scale technicolor
LXe	liquid Xenon
MACHO	massive compact halo object
MC	Monte Carlo
MD	matter dominance
MELA	matrix element likelihood approach
meV	milli-electronvolt
μ eV	micro-electronvolt
MeV	mega-electronvolt
MINOS	Main Injector Neutrino Oscillation Search
M/L	mass to luminosity ratio
MO	mineral oil
MS	main sequence star
MS	multisite
MS-DESI	Mid-Scale Dark Energy Spectroscopic Instrument

MSSM	minimum supersymmetric extension of the Standard Model
mSUGRA	*minimum supergravity*
MSW	*Mikheyev-Smirnov-Wolfenstein*
MW	mega watt
NC	neutral current
ND	nondegenerate
NF	neutrino factory
NGB	Nambu-Goldstone boson
NH	*normal hierarchy*
NIR	near infrared
NLO	next-to-leading order
NLSP	next to lightest superparticle
NNLO	next-to-next-to-leading order
NR	nonrelativistic
NS	normal spectrum
NOvA	NuMI Off-Axis Electron Neutrino Appearance
NUHM	Non-Universal Higgs Masses
NUMI	Neutrinos at the Main Injector
OAV	outer acrylic vessel
OD	outer detector
OPERA	Oscillation Project with Emulsion-tRacking Apparatus
OWS	outer water shield
P-AGB	post-asymptotic giant branch
PC	partially contained
PDF	probability density function
PMNS	Pontecorvo–Maki–Nakagawa–Sakata
PMTs	photomultiplier tubes
pNGBs	*pseudo Nambu-Goldstone bosons*
POT	protons on target
ppm	parts per million
PQ	Peccei-Quinn
PS	proton synchrotron
QBH	Quantum black hole
QCD	quantum chromodynamics
QD	quasi-degeneracy
QED	quantum electrodynamics
RD	radiation dominance
RGB	red-giant branch
RGE	renormalization group equation
RH	right-handed
RS	Randall-Sundrum
RSD	Red Shift-Space Distortions
SB	super-beam

SB	symmetry breaking
SCT	SemiConductor Tracker
SDSS	Sloan Digital Sky Survey
SGB	subgiant branch
SHM	Standard Halo Model
SI	spin independent
SM	Standard Model
SMA	small mixing angle
S/N	signal-to-noise ratio
SNAP	supernova acceleration probe
SNe	supernovae
SNIa	type Ia supernovae
SNO	Sudbury Neutrino Observatory
SNU	solar neutrino unit
SPS	Super-Proton-Synchrotron
SPT	South Pole Telescope
SQUID	Superconducting QUantum Interference Devices
SS	single-site
SSM	sequential standard model
SSM	standard solar model
SSV	stainless steel vessel
sUED	split UED
SUGRA	super-gravity
SuperK	SuperKamiokande
SUSY	super-symmetry
SW	Sachs–Wolfe
T2K	Tokai to Kamioka
TC	technicolor
TeV	tera-electronvolt
TO	turnoff
TPC	time projection chamber
UED	universal extra dimension model
V–A	vector–axial vector
VEV	vacuum expectation value
W.E.	water equivalent
WFIRST	Wide Field Infrared Survey Telescope
WIMP	weakly interacting massive particle
WL	weak lensing
WMAP	Wilkinson Microwave Anisotropy Probe
Xe-LS	xenon-loaded liquid scintillator

1
Higgs

1.1
Introduction

The Standard Model (SM) is a theory to explain interactions of elementary particles at the most fundamental level. Its essence can be summarized as follows.

1) The fundamental constituent blocks of matter are quarks and leptons.
2) The mathematical framework for the force dynamics are gauge theories.
3) The vacuum is in a kind of super-conducting phase.

The phase transition to the pseudo-superconducting phase is the action of the Higgs field. However, the precise dynamics of the symmetry breaking is not known. For this reason, the SM was believed to lose its predictive power on phenomena at energy scale over O(TeV).

The electroweak (EW) interaction of fermions (i.e., quarks and leptons) is mediated by the gauge particles $W^{\pm}$, Z^0. The role of the Higgs is to attach masses to the gauge particles and fermions without breaking the gauge symmetry. The symmetry is broken spontaneously by self-interactions of the Higgs field, which is referred to as the *Higgs mechanism*. It is the fundamental framework of the SM, and its basic notion is well founded because the SM reproduces experiments quite well.

However, it is also the least known sector in the SM. The Higgs potential was chosen for its simplicity and may not be realized in the real world. In fact, evidences are accumulating that some kind of symmetry is at work in the Higgs sector that is not yet identified. Only direct production of the Higgs particle and detailed investigation of its dynamical properties will clarify the properties of the Higgs field.

The gauge sector of the SM has been proved to reproduce experiments to a high degree of precision. From its perspective, the Higgs sector's role is only to provide masses to otherwise massless particles, and nothing else matters. Very little information on the Higgs is obtained from the gauge sector. Besides, there is no guiding principle to unify the mass generation mechanism of the fundamental

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

fermions (aka the Yukawa interaction), and this remains the weak point of the SM to qualify it as a unified theory. If a defect exists in the so-far-infallible SM, it is highly probable that it is found in the Higgs sector. As the Higgs particle was discovered at the large hadron collider (LHC) in 2012,1) the next urgent problem of the SM is to elucidate its dynamical structure.

In this chapter, we will discuss the fundamental properties of the Higgs particle and theoretical constraints on its mass, as well as how to proceed after its discovery [3–5]. We review methods to discover the Higgs, because properties of the Higgs are best elucidated by its detection strategy. We will also discuss the likely possibilities of the Higgs properties that go beyond the SM, including the supersymmetric extension and the possibility of strong dynamical breaking of the EW symmetry.

1.2
Higgs Interactions

1.2.1
Standard Model

The symmetry of the EW interaction in the SM is based on a mixture of $SU(2)$ and $U(1)$. The weak force in its original form, that is, before mixing and spontaneous symmetry breakdown, has chiral symmetry.

In $SU(2)$ terminology, the weak force carriers constitute an isospin triplet. All the left-handed fermions constitute doublets. All the right-handed particles belong to $SU(2)$ singlets ($I = I_3 = 0$), that is, they do not carry weak charges. In the SM, all the leptons can be classified by their isospin component as

$$\begin{array}{lll} \text{Leptons} & \begin{cases} I_3 = +\frac{1}{2} \\ I_3 = -\frac{1}{2} \end{cases} & \Psi_L = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L, \quad \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_L, \quad \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}_L \\ & I = I_3 = 0 & e_R^-, \ \mu_R^-, \ \tau_R^- \end{array} \tag{1.1}$$

The leptons that have $I_3 = 1/2$, that is, the neutrinos, are electrically neutral and those that have $I_3 = -\frac{1}{2}$ have electric charge $Q = -1$ in units of the positron charge.

---

1) The LHC is located at CERN in Geneva, and is designed to reach the total center-of-mass energy $\sqrt{s} = 14$ TeV with luminosity 5×10^{34} cm^{-2} S^{-1}. The detectors comprise two general-purpose detectors (A Toroidal LHC ApparatuS, ATLAS, and compact muon solenoid, CMS) and two dedicated detectors, LHCb for B-physics and ALICE for heavy-ion physics.

In the SM, right-handed neutrinos do not exist.[2] For the quarks

$$\text{Quarks} \quad \begin{cases} I_3 = +\frac{1}{2} \\ I_3 = -\frac{1}{2} \end{cases} \quad \begin{pmatrix} u \\ d' \end{pmatrix}_L \quad \begin{pmatrix} c \\ s' \end{pmatrix}_L \quad \begin{pmatrix} t \\ b' \end{pmatrix}_L \qquad (1.2)$$
$$I = I_3 = 0 \qquad u_R,\ d_R,\ c_R,\ s_R,\ t_R,\ b_R$$

where $D'^T \equiv (d', s', b')^T$ are the Cabibbo–Kobayashi–Maskawa (CKM) rotated fields:

$$\begin{bmatrix} d' \\ s' \\ b' \end{bmatrix} = V_{\text{CKM}} \begin{bmatrix} d \\ s \\ b \end{bmatrix}, \qquad V_{\text{CKM}} = \begin{bmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{bmatrix} \qquad (1.3)$$

The quarks with $I_3 = 1/2$ have $Q = 2/3$, and those with $I_3 = -1/2$ have $Q = -1/3$. Each quark carries another degree of freedom, that is, three colors, which are the source of the strong interaction. Its dynamics constitutes a field of its own, but in the discussion of the EW force, we put aside their interactions and simply consider the fact that they only provide three extra degrees of freedom.

We denote the Higgs doublet[3] the left-handed electron doublet, and the gauge boson triplet as

$$\Phi = \begin{bmatrix} \phi^+ \\ \phi^0 \end{bmatrix}, \qquad \Psi_L = \begin{bmatrix} \nu_L \\ e_L^- \end{bmatrix}, \qquad \boldsymbol{W}_\mu = (W_\mu^+,\ W_\mu^0,\ W_\mu^-) \qquad (1.4)$$

The original Lagrangian of the EW interaction before mixing and spontaneous symmetry breaking is given by

$$\begin{aligned} \mathscr{L}_{\text{EW}} &= \overline{\Psi} i\gamma^\mu D_\mu \Psi - \frac{1}{4}\boldsymbol{F}_{\mu\nu} \cdot \boldsymbol{F}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \\ &\quad + (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi) - \gamma_e [\overline{e}_R (\Phi^\dagger \Psi_L) + (\overline{\Psi}_L \Phi) e_R] \end{aligned} \qquad (1.5)$$

$$\boldsymbol{F}_{\mu\nu} = \partial_\mu \boldsymbol{W}_\nu - \partial_\nu \boldsymbol{W}_\mu - g_W \boldsymbol{W}_\mu \times \boldsymbol{W}_\nu \qquad (1.6a)$$
$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \qquad (1.6b)$$

---

[2] In reality, they do exist as demonstrated by the discovery of the neutrino oscillation. In the context of this textbook, no inconvenience is encountered by assuming the massless neutrino in this chapter and it greatly simplifies discussions. The neutrino oscillation phenomena will be treated in the next chapter.

[3] We loosely call Φ or all the four scalar fields as the Higgs. But to be more accurate, after the symmetry breakdown, ϕ^+ and the imaginary part of ϕ^0 are absorbed by the gauge bosons to give them mass, and only one field, the real part of ϕ^0, remains as a massive physical field. It is renamed as h, which is the genuine Higgs field.

$$D_\mu = \partial_\mu + ig_W \boldsymbol{W}_\mu \cdot \boldsymbol{t} + i(g_B/2)Y \tag{1.6c}$$

$$V(\Phi) = \lambda\left(|\Phi|^2 + \frac{\mu^2}{2\lambda}\right)^2 \qquad \lambda > 0 \tag{1.6d}$$

where $\boldsymbol{t}$ is the generator of the $SU(2)$ symmetry group. Notice that, except in the Higgs potential, there are no mass terms (quadratic term in the fields) in the Lagrangian.[4]

We shall use ν_e, e^-, and so on, to denote the quantized fields, that is, $\nu_e(x) = \psi_{\nu_e}(x)$, $e^-(x) = \psi_e(x)$, and so on, where there is no confusion. Here, we have written down only the Lagrangian of $\Psi^T = (\nu_e, e^-)$, which will be needed in the following discussions. The Lagrangian for other fermions can be written down similarly.

The first line of Eq. (1.5) is referred to as the *gauge sector* and the second line as the *Higgs sector*. $V(\Phi)$ is the self-interacting potential of the Higgs field. The whole expression satisfies the $SU(2) \times U(1)$ gauge symmetry manifestly. It is important to remember that both the gauge and the Higgs sectors are constructed to respect the gauge symmetry separately. The last term of Eq. (1.5), referred to as the *Yukawa interaction*, was added to generate fermion masses. It can be written down as

$$\overline{e}_R(\Phi^\dagger \Psi_L) + (\overline{\Psi}_L \Phi)e_R = \overline{e}_R \nu_{eL}\phi^- + \overline{\nu}_{eL} e_R \phi^+ + \overline{e}_R e_L \phi^{0^\dagger} + \overline{e}_L e_R \phi^0 \tag{1.7}$$

The self-interaction of the Higgs is the cause of the spontaneous symmetry breakdown of the $SU(2)_L \times U(1)$, giving mass to the gauge bosons and the fermions.

Electroweak mixing: As the neutral component of $\boldsymbol{W}$ and B couple to the same fermions, mixing occurs, and physical neutral gauge bosons γ and Z^0 are expressed as

$$\begin{aligned} Z_\mu &= \frac{1}{\sqrt{g_W^2 + g_B^2}}(g_W W_\mu^0 - g_B B_\mu) \equiv \cos\theta_W W_\mu^0 - \sin\theta_W B_\mu \\ \gamma_\mu &= \frac{1}{\sqrt{g_W^2 + g_B^2}}(g_B W_\mu^0 + g_W B_\mu) \equiv \sin\theta_W W_\mu^0 + \cos\theta_W B_\mu \end{aligned} \tag{1.8}$$

where θ_W (or $\sin\theta_W$) is referred to as the *Weinberg angle*. As a result, the gauge interaction of the fermions is described by the interaction Lagrangian

$$\begin{aligned} -\mathscr{L}_{\text{int}} &= \frac{g_W}{2}\overline{\Psi}_L \gamma^\mu \boldsymbol{W}_\mu \cdot \boldsymbol{\tau}\Psi_L + \frac{g_B}{2}\Psi\gamma^\mu B_\mu \Psi \\ &= \frac{g_W}{\sqrt{2}}\overline{\Psi}_L \gamma^\mu \left(W_\mu^+ \tau_+ + W_\mu^- \tau_-\right)\Psi_L \\ &\quad + g_Z \overline{\Psi}\gamma^\mu (I_{3L} - Q\sin^2\theta_W)\Psi Z_\mu + e\overline{\Psi}\gamma^\mu Q\Psi A_\mu \end{aligned} \tag{1.9}$$

---

[4] The quadratic as well as the quartic terms exist in the SM Higgs potential. But in some extensions of the SM, they are absent in the tree Lagrangian. They are generated radiatively as a result of the Higgs–gauge interaction [6].

where $\tau_\pm = (\tau_1 \pm \tau_2)/2$ and the coupling strength of Z and A is given by

$$g_Z = \frac{g_W}{\cos\theta_W} = \frac{e}{\sin\theta_W \cos\theta_W}, \qquad e = g_W \sin\theta_W \tag{1.10}$$

1.2.2
Lagrangian After Symmetry Breaking

The whole Lagrangian of the SM satisfies manifestly the $SU(2) \times U(1)$ gauge symmetry. Symmetry breaking occurs when the coefficient of the quadratic term in the potential is driven negative. It happens when the environmental temperature goes below a certain critical temperature. The potential minimum moves away from $\Phi = 0$, that is, the vacuum moves to where one of its components is finite.[5]

We say that the Higgs field has acquired the vacuum expectation value (VEV) $<\phi^0> = v/\sqrt{2}$. The factor $1/\sqrt{2}$ is conventional. The vacuum is no longer at $\Phi = 0$. As the symmetry around two different points are different, that is, the potential has no original symmetry at the new vacuum point, we say the symmetry is broken. In the terminology of condensed matter physics, the ground state is now Bose condensate, with v representing an order parameter. Physical phenomena that are observed as excitations from the new vacuum no longer exhibit the original symmetry. Notice, however, that the symmetry of the potential is not really lost. Physical phenomena are generally small excitations around the vacuum, which is a local minimum of the potential. Power expansions of the potential around $\Phi = 0$ and around $\phi^0 = v/\sqrt{2}$ have different mathematical expression. Naturally, a physical phenomenon interpreted from viewpoint of the new vacuum look different from that in the old vacuum. Mathematically, it is possible to work using the field variables based on the old vacuum compromising easy physical interpretation. In this sense, the symmetry is not broken but hidden.

The condition that $\phi^0 = v/\sqrt{2}$ is at potential minimum sets the value of μ^2.

$$\left.\frac{\partial V}{\partial \phi^0}\right|_{\phi^0 = v/\sqrt{2}} = 0 \quad \xrightarrow{\text{SB}} \quad \mu^2 = -\lambda v^2 \tag{1.11}$$

where SB denotes the symmetry breaking. Without loss of generality, the Higgs field before and after the symmetry breakdown can be reexpressed as follows:

$$\Phi = e^{i\frac{\boldsymbol{\omega}}{v}\cdot\frac{\boldsymbol{\tau}}{2}} \begin{bmatrix} 0 \\ \phi^0 \end{bmatrix} \equiv U^{-1} \begin{bmatrix} 0 \\ \phi^0 \end{bmatrix} \quad \xrightarrow{\text{SB}} \quad \Phi' = U\Phi = \begin{bmatrix} 0 \\ \frac{v+h}{\sqrt{2}} \end{bmatrix} \tag{1.12}$$

"h" is the physical Higgs field we are interested in and is most significant in this chapter. The Higgs components $\boldsymbol{\omega}$ in the phase are removed by the gauge transformation U and become the third component of the massive gauge bosons. In mathematical language, the $\boldsymbol{\omega}$ fields are gauged away. Stated differently, the spontaneous symmetry breakdown is equivalent to choosing a gauge and fixing it. We write the covariant derivative and the Higgs doublet after the symmetry breaking as

---

[5] It is customary to choose the vacuum point at $\phi^0 = \frac{v}{\sqrt{2}}$, $\mathrm{Im}[\phi^0] = \phi^+ = \phi^- = 0$.

$$D_\mu\Phi \xrightarrow{SB} UD_\mu U^{-1}U\Phi \equiv D'_\mu\Phi' = \left[\partial_\mu + ig_W\mathbf{W}'_\mu \cdot \boldsymbol{t} + i\frac{g_B}{2}B_\mu\right]\begin{bmatrix} 0 \\ \dfrac{v+h}{\sqrt{2}} \end{bmatrix} \tag{1.13}$$

Then we rename $D'_\mu\Phi'$, $\mathbf{W}'_\mu$ as $D_\mu\Phi$, $\mathbf{W}_\mu$.

In terms of the newly defined field variables, the Higgs sector of the Standard Model Lagrangian is reexpressed as

$$\begin{aligned}\mathscr{L}_h =& (D_\mu\Phi)^\dagger(D^\mu\Phi) - \lambda\left(\Phi^\dagger\Phi - \frac{v^2}{2}\right)^2 - y_e[\bar{e}_R(\Phi^\dagger\Psi_L) + (\overline{\Psi}_L\Phi)e_R]\\ &\xrightarrow{SB} \frac{1}{2}\left(\partial_\mu h\partial^\mu h - m_h^2h^2\right) - \left(\lambda v h^3 + \frac{\lambda}{4}h^4\right)\\ &+ 2(\sqrt{2}G_F)^{\frac{1}{2}}h\left(m_W^2W_\mu^+W^{-\mu} + \frac{m_Z^2}{2}Z_\mu Z^\mu\right)\\ &+ \sqrt{2}G_Fh^2\left(m_W^2W_\mu^+W^{-\mu} + \frac{m_Z^2}{2}Z_\mu Z^\mu\right)\\ &+ \sum_f\left(m_f + \sqrt{2}G_Fm_fh\right)\bar{f}f\end{aligned} \tag{1.14}$$

where the electron (e) term has been expanded to include all the fermions (f). The coupling constants G_F, y_f, and λ are related to the masses by

$$g_W = \frac{e}{\sin\theta_W} = 2(\sqrt{2}G_F)^{\frac{1}{2}}m_W \tag{1.15a}$$

$$g_Z = \frac{e}{\sin\theta_W\cos\theta_W} = 2(\sqrt{2}G_F)^{\frac{1}{2}}m_Z \tag{1.15b}$$

$$\frac{1}{v} = \frac{g_W}{2m_W} = (\sqrt{2}G_F)^{\frac{1}{2}} = \frac{1}{246\ \text{GeV}} \tag{1.15c}$$

$$m_h^2 = 2\lambda v^2, \qquad m_f = \frac{y_fv}{\sqrt{2}} \tag{1.15d}$$

Notice that the mass of the fermions is directly proportional to the Yukawa coupling constant. After the symmetry breaking, the SM has mass terms for the gauge bosons, fermions, and the Higgs itself. As one can see from the above expressions, the coupling strength of the Higgs is directly proportional to the mass of the particles to which it couples. The self-coupling strength λ is also proportional to the mass of the Higgs (squared). λ is an unknown parameter, hence the Higgs mass cannot be determined a priori. The Lagrangian Eq. (1.14) is the starting point of the Higgs interactions. Feynman rules for the Higgs interaction are given in Figure 1.1.

1.2.3
Decay Modes

Decays to a Fermion Pair Once the Lagrangian is given, it is straightforward to write down the matrix element for the decay of the Higgs to two fermions in the

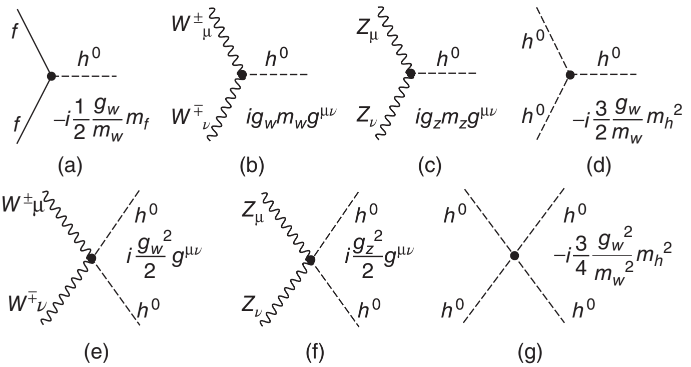

f
h^0
f
$-i\frac{1}{2}\frac{g_w}{m_w}m_f$
(a)
$W^{\pm}_{\mu}$
h^0
$W^{\mp}_{\nu}$
$ig_w m_w g^{\mu\nu}$
(b)
Z_{μ}
h^0
Z_{ν}
$ig_z m_z g^{\mu\nu}$
(c)
h^0
h^0
h^0
$-i\frac{3}{2}\frac{g_w}{m_w}m_h{}^2$
(d)
$W^{\pm}_{\mu}$
h^0
$i\frac{g_w{}^2}{2}g^{\mu\nu}$
$W^{\mp}_{\nu}$
h^0
(e)
Z_{μ}
h^0
$i\frac{g_z{}^2}{2}g^{\mu\nu}$
Z_{ν}
h^0
(f)
h^0
h^0
$-i\frac{3}{4}\frac{g_w{}^2}{m_w{}^2}m_h{}^2$
h^0
h^0
(g)

Figure 1.1 Feynman rules for the Higgs interaction. The coupling constants γ_f, λ, and so on, are reexpressed in terms of the mass of the particles to which the Higgs couples.

tree approximation.

$$\mathcal{M}(h \to f\bar{f}) = (\sqrt{2}G_F)^{\frac{1}{2}} m_f \bar{u}(p_1)v(p_2) \tag{1.16}$$

The decay width becomes

$$\Gamma(h \to f\bar{f}) = \frac{N_c G_F m_f^2 m_h}{4\sqrt{2}\pi}\left(1 - \frac{4m_f^2}{m_h^2}\right)^{3/2} \tag{1.17}$$

where $N_c = 1$ for leptons and $N_c = 3$ for quarks. As the coupling strength is proportional to the mass, the dominant mode is the decay to the heaviest fermions that is allowed energetically. For $m_h < 2m_W$, $h \to b\bar{b}$ is the most dominant decay mode.

Decay to Bosons If the Higgs mass is greater than $2m_W$ or $2m_Z$, it can decay to W or Z pairs. The decay amplitude to the W pair can be calculated from the third line of Eq. (1.14) to give

$$\mathcal{M}(h \to W^+W^-) = 2(\sqrt{2}G_F)^{\frac{1}{2}} m_W^2 \varepsilon_\mu^*(p_1, \lambda_1)\varepsilon^\mu(p_2, \lambda_2) \tag{1.18}$$

where $p_1, p_2, \lambda_i, \varepsilon^\mu$ denote the momenta, polarization state, and polarization vectors of $W^{\pm}$. Then the decay width to pairs with transverse ($\lambda = \pm$) and longitudinal

($\lambda = 3$) polarization are given by

$$\Gamma(h \to W_T W_T) = \frac{G_F m_W^4}{2\sqrt{2}\pi m_h}(1 - 4\gamma_W^2)^{1/2} \tag{1.19a}$$

$$\Gamma(h \to W_L W_L) = \frac{G_F m_h^3}{8\sqrt{2}\pi}(1 - 2\gamma_W^2)^2(1 - 4\gamma_W^2)^{\frac{1}{2}} \tag{1.19b}$$

$$\Gamma(h \to W^+ W^-) = 2\Gamma(h \to W_T W_T) + \Gamma(h \to W_L W_L) \tag{1.19c}$$

$$= \frac{G_F m_h^3}{8\sqrt{2}\pi}(1 - 4\gamma_W^2 + 12\gamma_W^4)(1 - 4\gamma_W^2)^{\frac{1}{2}} \tag{1.19d}$$

$$\text{where} \quad \gamma_W = \frac{m_W}{m_h}, \quad \gamma_Z = \frac{m_Z}{m_h} \tag{1.19e}$$

From the above expression, one sees that, if $m_h \gg m_W$, the W pairs that decayed from the Higgs are dominantly polarized longitudinally. The decay width to Z can be obtained similarly.

$$\Gamma(h \to ZZ) = \frac{G_F m_h^3}{16\sqrt{2}\pi}(1 - 4\gamma_Z^2 + 12\gamma_Z^4)(1 - 4\gamma_Z^2)^{1/2} \tag{1.20}$$

There is a factor 2 difference between the W pair decay modes which originate from the Bose statistics for identical particles. For $m_h \gg m_Z, \Gamma(h \to WW) \approx 2\Gamma(h \to ZZ)$.

Carrying out calculations for all possible decay modes and adding all, one can obtain the total decay width of the Higgs particle. If $m_h < 2m_W$, the main decay mode is $h \to b\bar{b}$. If $m_h > 2m_W$, it mainly decays to WW and ZZ. Figure 1.2 shows the total decay width and branching ratios of the Higgs as a function of the Higgs mass.

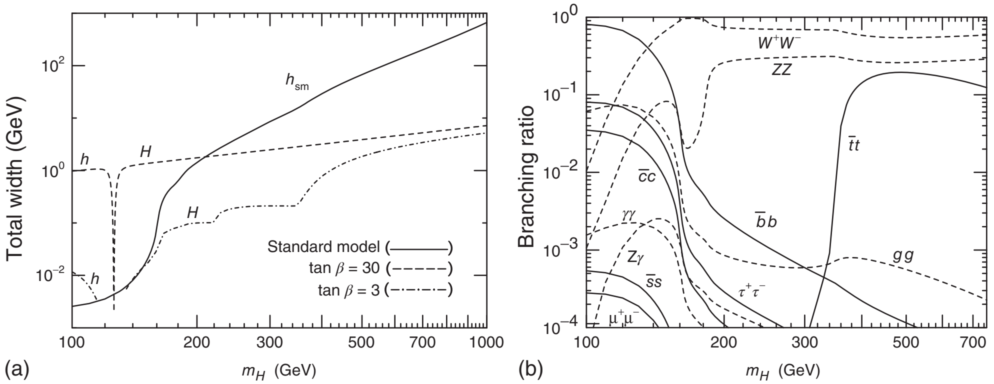


Figure 1.2 (a) Total decay width of the Standard Model Higgs. Also shown are those of two neutral Higgs (h and H) for the supersymmetric extension of the Higgs sector. $\tan\beta = v_2/v_1$ is the ratio of the two vacuum expectation values of the condensed Higgs field of the supersymmetry. (b) Branching ratios of the dominant decay modes of the Standard Model Higgs boson. (Reproduced with permission of [7].)

The reason why $\Gamma(h \to ZZ)$ decreases around $m_h \sim 160$ GeV is because the channel $h \to WW$ opens here.

1.3
Mass

The Higgs was discovered in 2012 with mass $m_h \simeq 125$ GeV. However, its mass value is not just one parameter among many. The size of the Higgs mass has an important significance in considering the mechanism of the spontaneous symmetry breakdown and the future scenario of the physics beyond the SM.

1.3.1
Predictions from EW Data

Much information had been obtained from radiative correction data at the large electron–positron (LEP) collider and the Tevatron. As we learned in Chapter 5 of [2], main contributions to the radiative corrections come from heavy particles, namely top and the Higgs. From the Z decay and W production data, one can determine the value of the Weinberg angle $\sin\theta_W$ and the coupling ratio ρ of the neutral versus charged current interaction.

$$\begin{aligned}
&\rho \equiv \rho_0(1+\Delta\rho) \equiv \frac{G_N}{G_W}{}^{6)} = \frac{m_W^2}{m_Z^2\cos\theta_W^2}\\
&\Delta\rho = \delta\rho_t + \delta\rho_h + \cdots\\
&\delta\rho_t \simeq \frac{3G_F m_t^2}{8\sqrt{2}\pi^2} \approx 0.0096\left(\frac{m_t}{173\,\text{GeV}}\right)^2\\
&\delta\rho_h \simeq -\frac{3G_F m_Z^2\sin\theta_W^2}{8\sqrt{2}\pi^2}\left(\ln\frac{m_h^2}{m_W^2}-\frac{5}{6}\right)^2
\end{aligned} \tag{1.21}$$

As the top mass was determined by direct productions at the Tevatron, its value can be used to test the validity of the EW prediction. The observed top mass agreed quite well with the prediction given by the precision EW data [8]. It is a major triumph of the SM. Now that the top mass is given, Eq. (1.21) can be used to predict the Higgs mass. As the Higgs contribution is logarithmic, sensitivity to the Higgs mass is weak. An overall fit to the world data for the mass value prior to LHC operation is given in Figure 1.3.

In this way, one can predict the value of the SM Higgs boson mass. A very recent analysis shows [10, 12]

$$m_h = 120^{+12}_{-05}\ \text{GeV} \tag{1.22}$$

---

6) G_N is the four-Fermi coupling constant of the neutral current interaction corresponding to G_F of the charged current. At the tree level, $G_N = G_F$. Note that G_N is also used to denote the Newton's gravitational constant elsewhere.

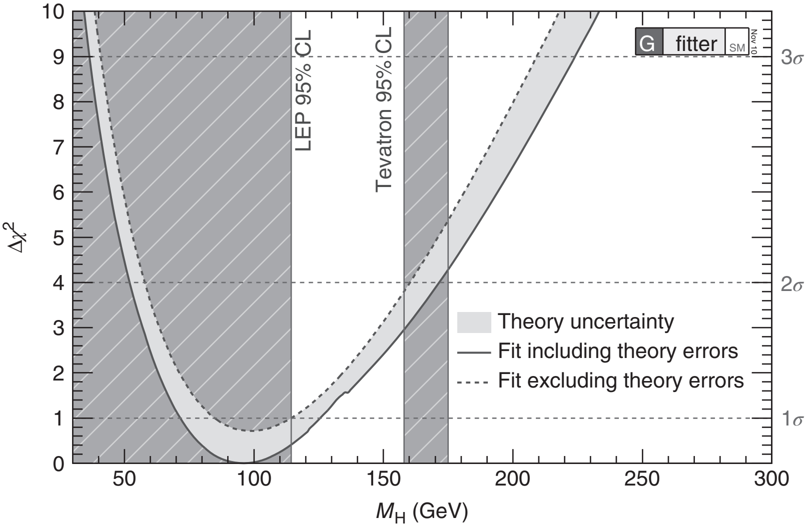


Figure 1.3 Higgs mass exclusion plot as a function of the Higgs mass before the Higgs discovery. (Reproduced with permission of [9–12].)

leading to a 95% CL limit in the SM

$$114 < m_h < 144 \ \text{GeV} \tag{1.23}$$

Here we discuss the theoretical implications of this value.

1.3.2
Vacuum stability

Comparison of theoretical radiative corrections with precision experimental data predicts a rather low Higgs mass. Let us consider first what theoretical constraints can be obtained from the SM framework. A constraint for the lower mass limit can be obtained from vacuum stability. One condition for spontaneous symmetry breaking of the vacuum was that λ, that is, the coefficient of the quartic term in the Higgs potential, must remain positive.

$$\begin{aligned} V(\phi) &= \lambda\left(\Phi^\dagger\Phi - \frac{v^2}{2}\right)^2 = V(0) - \frac{m_h^2}{2}\Phi^\dagger\Phi + \lambda(\Phi^\dagger\Phi)^2 \\ m_h^2 &= 2\lambda v^2 \end{aligned} \tag{1.24}$$

A negative value of λ makes the vacuum unstable for large values of $|\Phi|$. Positivity of λ is guaranteed only at the tree level. Taking into account the radiative corrections to the Higgs propagator, the value of the coupling constant λ is governed at one-loop level by the following renormalization group equation [3, 13].[7]

$$\frac{d\lambda}{d\tau} = \beta(\lambda) = \frac{3}{4\pi^2}(\lambda - \lambda_+)(\lambda - \lambda_-), \qquad \lambda_- \leq \lambda_+ \tag{1.25a}$$

$$\beta(\lambda) = \frac{1}{16\pi^2}\left[12\lambda^2 + 6\lambda y_t^2 - 3y_t^4 - \frac{3}{2}\lambda(g_W^2 + g_B^2) + \frac{3}{16}\left\{2g_W^4 + (g_W^2 + g_B^2)^2\right\}\right] \tag{1.25b}$$

where $\tau = \ln(Q^2/Q_0^2)$. $y_t = \sqrt{2}m_t/v$ is the Yukawa coupling constant, and g_W and g_B are the $SU(2)$ and $U(1)$ gauge coupling constants, respectively. Contributions from quarks other than the top are neglected because the coupling strength is directly proportional to its mass value. The λ that obeys Eq. (1.25) is no longer a constant, but a dynamical variable that depends on τ. The original λ can be considered as that defined by the condition $\lambda = \lambda(Q)|_{Q=Q_0\,(=v)}$. Inclusion of the radiative corrections changes the potential shape, and hence the VEV of the Higgs ϕ^0 has to be redefined as that which gives the minimum to the modified potential.

Let us disregard, for simplicity, the Q^2 dependence of y_t, g_W, g_B. Then, $\beta(\lambda)$ is a quadratic function of λ and has zeros at $\lambda = \lambda_\pm$. For $4m_t^4 > 2m_W^4 + m_Z^4$, we have $\lambda_- < 0 < \lambda_+$. The solution depends on the magnitude of the initial value $\lambda_0 = \lambda(Q_0)$. Depending on whether $\lambda_0 > \lambda_+$ or $0 < \lambda_0 < \lambda_+$, the solution to Eq. (1.25) is given by

$$\frac{\lambda - \lambda_-}{\lambda - \lambda_+} = \pm A e^{-\delta\tau} \tag{1.26a}$$

$$\delta = \frac{3}{4\pi^2}(\lambda_+ - \lambda_-) \tag{1.26b}$$

For $\lambda_0 > \lambda_+$, the value of λ will always remain positive. If $0 < \lambda_0 < \lambda_+$, the beta function is negative ($\beta(\lambda) < 0$). In this case, for sufficiently large Q, λ becomes negative at a certain value $Q = \Lambda$. The resultant potential is negative at large ϕ and has no minimum. In other words, the vacuum is unstable. This happens for small λ, in which case the top quark contribution dominates and drives λ to a negative value for sufficiently large Q^2. For small λ, Eq. (1.25) becomes

$$\frac{d\lambda}{d\tau} \simeq \frac{1}{16\pi^2}\left[-3y_t^4 + \frac{3}{16}\left\{2g_W^4 + (g_W^2 + g_B^2)^2\right\}\right] \tag{1.27}$$

To provide an intuitive understanding through easy analytic implementation, we assume y_t, g_W, and g_B are constant, and integrate Eq. (1.27) and obtain

$$\lambda(\Lambda) = \lambda(v) + \frac{1}{16\pi^2}\left[-3y_t^4 + \frac{3}{16}\left\{2g_W^4 + (g_W^2 + g_B^2)^2\right\}\right]\ln\left(\frac{\Lambda^2}{v^2}\right) \tag{1.28}$$

To ensure that $\lambda(\Lambda)$ remains positive, the Higgs mass ($= 2\lambda v^2$) must satisfy

$$m_h^2 > \frac{v^2}{8\pi^2}\left[3y_t^4 - \frac{3}{16}\left\{2g_W^4 + (g_W^2 + g_B^2)^2\right\}\right]\ln\left(\frac{\Lambda^2}{v^2}\right) \tag{1.29}$$

More elaborate calculation can be carried out by integrating Eq. (1.25). A detailed calculation incorporating two loop-level radiative corrections has been carried out in [14] and gives the lower limit as a function of Λ, which is shown as the lower curve of Figure 1.4.

---

[7] For basics of the renormalization group equation, see Section 7.2 of [2].

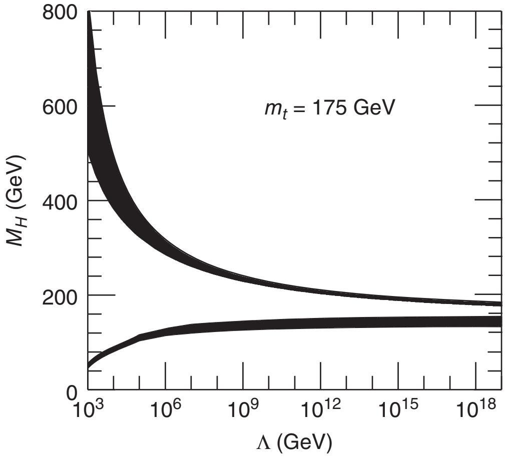


Figure 1.4 Triviality (upper) bound and vacuum stability (lower) bound on the Higgs boson mass as functions of the new physics or cutoff scale Λ for a top quark mass $m_t = 175 \pm 6$ GeV and $\alpha_s(mz) = 0.118 \pm 0.002$. The allowed region lies between the bands, and the colored/shaded bands illustrate the impact of various uncertainties. (Reproduced with permission of [13–15].)

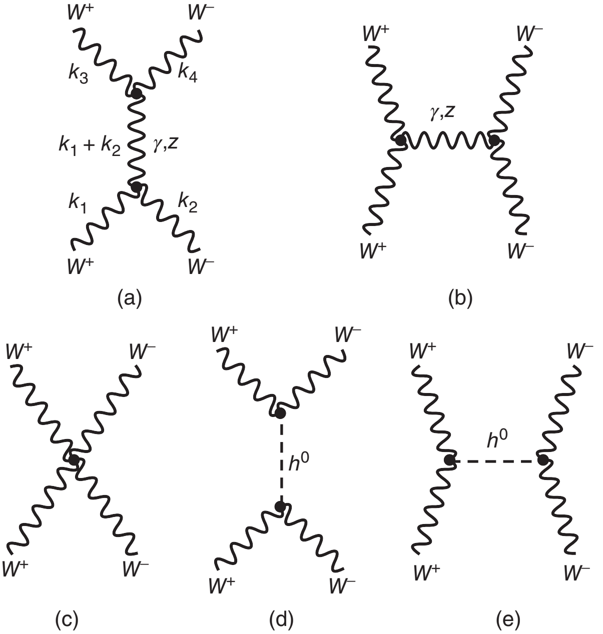
A state-of-the-art quantum correction at the next-to-next-to-leading order (NNLO) calculations has been carried out recently [16]. Assuming the validity of Eq. (1.25) all the way up to the grand unified theory (GUT) energy, one obtains

$$m_h > 129.4 + 1.4 \times \left(\frac{m_t\,(\text{GeV}) - 173.1}{0.7} \right) - 0.5 \times \left(\frac{\alpha_s(m_Z) - 0.1184}{0.0007} \right) \pm 1.0\ \text{GeV} \tag{1.30}$$

1.3.3
Theoretical Upper Limit

Perturbative Unitarity The Higgs mass cannot be indefinitely large. The simplest argument is supplied by the unitarity condition of the tree-level scattering amplitude. Let us consider scattering of the gauge boson W. Feynman diagrams that contribute to the elastic scattering are presented in Figure 1.5.

The gauge invariance provides a compensating mechanism for particular amplitudes from diverging to infinity and suppresses divergence to at most the logarithmic level (see arguments in Chapter I of [2]). For instance, each amplitude in Figure 1.5a–c grows $\sim s^2$ but combined together they cancel each other and the divergence is at most $\sim s$. Terms that grow by $\sim s$ are cancelled by the Higgs intermediate processes (Figure 1.5d,e), and the resultant overall divergence is at most $\sim \ln s$. At the tree level, the contribution of the Higgs intermediate scattering amplitude is approximately given by [17]



(a) (b) (c) (d) (e)

Figure 1.5 Feynman diagrams for $W^+W^- \to W^+W^-$.

$$\mathscr{M}(W_L^+ W_L^- \to W_L^+ W_L^-) = -\sqrt{2} G_F m_h^2 \left[\frac{s}{s - m_h^2} + \frac{t}{t - m_h^2} \right] + O\left(\frac{m_W^2}{s} \right) \tag{1.31}$$

We have considered only the scattering amplitude by longitudinally polarized Ws because they are the ones that give bad divergence at large s. This is due to the fact that the longitudinal polarization produces a factor $\sim p_\mu p_\nu / m_W^2$ in the denominator of the gauge boson propagator. Decomposing the scattering amplitude into partial waves, it can be expressed as

$$\mathscr{M} = 16\pi \sum (2J + 1) a_J(s) P_J(\cos\theta) \tag{1.32}$$

At high energy ($s \gg m_W^2$), the S wave amplitude can be extracted to give

$$a_0(W_L W_L \to W_L W_L) = -\frac{G_F m_h^2}{8\sqrt{2}\pi} \left[2 + \frac{m_h^2}{s - m_h^2} - \frac{m_h^2}{s} \ln\left(1 + \frac{s}{m_h^2} \right) \right] \tag{1.33}$$

At sufficiently high energy, the first term is dominant. The unitarity requires $|a_0| \leq 1$. As the Born amplitude is real, the requirement becomes $|\mathrm{Re} A_0| \leq 1/2$. Then

$$m_h \leq \left[\frac{2\sqrt{2}\pi}{G_F} \right]^{\frac{1}{2}} = 872\,\mathrm{GeV} \tag{1.34}$$

Another argument one may use is that, for the Higgs to qualify as an elementary particle, its mass has to exceed its width, that is, $m_h > \Gamma_h$. This gives the condition

$$\Gamma(h \to WW + ZZ) \simeq \frac{3G_F m_h^3}{16\sqrt{2}\pi} < m_h \tag{1.35}$$

and $m_h < 1.4$ TeV is obtained. However, if the Higgs mass is as large as this, it means λ is also very large ($\lambda/4\pi \geq O(1)$) because $m_h^2 = 2\lambda v^2$. The self-interaction of the Higgs is strong, and one questions the validity of the perturbative calculation. All one can claim is that the Higgs mass should be less than ~ 1 TeV.

Triviality Another constraint on the upper limit can be obtained from the renormalization group equation. As we are discussing a possible outcome at large λ, we can neglect all the terms except λ in Eq. (1.25). Then we can obtain the following solution. Setting $Q_0 = v$

$$\frac{1}{\lambda(v)} - \frac{1}{\lambda(Q)} = \frac{3}{4\pi^2} \ln\left(\frac{Q^2}{v^2}\right) \tag{1.36}$$

In order for the above equation to be valid in the perturbative approach, $\lambda(Q)$ has to be finite and reasonably small. Then, $\lambda(v)$ can only vanish in the limit $Q \to \infty$. This means that the equation of motion that the Higgs obeys must be a free equation without interaction. It is a "trivial" solution. Rewriting the above formula, one obtains

$$\lambda(Q) = \frac{\lambda(v)}{1 - \dfrac{3}{4\pi^2}\lambda(v) \ln\left(\dfrac{Q^2}{v^2}\right)} \tag{1.37}$$

For sufficiently large Q^2, the denominator vanishes, which is referred to as the *Landau pole*, and the value of λ diverges. As the perturbative treatment fails for large λ, it means that the equation that the Higgs field satisfies has to be considered as an effective theory valid only at low energy. The perturbative approach may be justified for $Q \lesssim \Lambda$, for which $\lambda(Q)$ remains finite and sufficiently small. The maximum value of $\lambda(v)$ within the constraint can be obtained from Eq. (1.36) by setting $\lambda(Q) = \infty$. Therefore

$$m_h^2 \leq 2v^2 \max[\lambda(v)] = \frac{8\pi^2 v^2}{3\ln(\Lambda^2/v^2)} \tag{1.38}$$

If we assume that the perturbative approach is valid and λ remains small until $\Lambda = M_{\text{Planck}} \simeq 10^{19}$ GeV, we obtain $m_h < 150$ GeV. This value does not depend on how we set the value of $\lambda(Q)$. If we use $\lambda(Q) = 1$ instead of $\lambda(Q) = \infty$, the numerical value of the Higgs mass is almost the same.

This is an interesting fact. Suppose there is a grand unification at large Q ($\sim 10^{16}$ GeV), and no new physics comes in until the grand unification scale, the Higgs has to be light. The supersymmetry (SUSY) approach we discuss later, in which the Higgs is considered as elementary, falls in this category. Conversely, if the Higgs is heavy, for which it could be a composite, the new physics will appear at relatively low energy. Assuming it happens actually, we may rephrase the existence condition of the Higgs as $m_h < \Lambda$, which leads to $m_h \lesssim 800$ GeV. The upper curve of Figure 1.4 shows how the upper limit of the Higgs mass changes as a function of the cutoff Λ. The allowed range of the Higgs mass depends on where we set the cutoff Λ. If we assume no new physics until the GUT energy ($\sim 10^{16}$ GeV), we have a severe constraint $140 \lesssim m_h \lesssim 200$ GeV. The mass of the discovered Higgs is less than the lower limit but at the fringe of the limit.

Till date, we argued assuming that a new physics will appear at large λ. Is this assumption correct? The large λ may simply mean breakdown of the perturbation theory and may not necessarily guarantee the appearance of new physics. If the perturbative approach fails for $Q > \lambda_{\text{NT}}$, the region $\Lambda_{\text{NT}} < Q < \Lambda$ is where no new physics appears but nonperturbative approach is necessary. Whether such a region exists can be probed using lattice quantum chromodynamics (LQCD). Here, the lattice interval plays the role of the cutoff. Within the lattice QCD formalism, one can perform similar calculations as we did using the perturbation theory, and an upper limit of the Higgs mass was obtained as $m_h < 640$ GeV [18,19]. It appears that there is no region where the nonperturbative treatment is required. The conclusion we derived using the perturbation theory seems reliable. Therefore, if the Higgs was not found with mass $m_h < 1$ TeV, we could expect new physics nearby. That is, the assumption $m_h \lesssim 1$ TeV seems reasonable. Note, however, $m_h \lesssim 1$ TeV is the theoretical maximum value within reasonable allowance and was not the best guess value. One should remember that predictions of the Higgs, be it phenomenological guess or renormalization group equation argument, pointed to a rather low Higgs mass.

Metastable Vacuum So far we assumed a positive λ, which is necessary for a stable vacuum. However, even if λ goes negative, as long as the vacuum life is longer than the cosmic life time, it is still a viable solution [20]. It is argued that, if the Higgs is light, the vacuum develops an instability below the Planck scale but the EW vacuum is still sufficiently long-lived [21]. The observed value of the Higgs mass ($m_h \sim 125$ GeV, see Section 1.7.6) at LHC is lower than that in Eq. (1.30), which was obtained assuming the vacuum stability within the framework of the SM extrapolated all the way to the Planck scale. Given the LHC mass value, the stability of the EW vacuum has been reevaluated, and the result is shown in Figure 1.6a [8,22]. Indeed, the quartic coupling constant λ goes negative at large scale, which points to the metastability or instability of the SM vacuum.

The LHC discovery of the Higgs mass ($m_h \simeq 125$ GeV) is at the verge of vacuum metastability. Figure 1.6b shows that the mass value of the Higgs ($m_h = 125 \pm 1$ GeV) points to a position in the metastable region [16,23,24]. The stability condition is sensitive also to the value of the top mass and the strong coupling constant α_s. Within the present accuracy ($\Delta m_h = \pm 1$ GeV and $\Delta m_t = \pm 2$ GeV), the allowed region extends both to the stable and the unstable region. A future experiment at the e^-e^+ linear collider (ILC) could reach $\Delta m_h = \pm 50$ MeV, and $\Delta m_t = \pm 200$ MeV, and narrow down the region to the small circle in Figure 1.6b [23].

In summary, the question of the Higgs mass value is an important one directly related to how the new physics will appear. The observation of a Higgs mass of

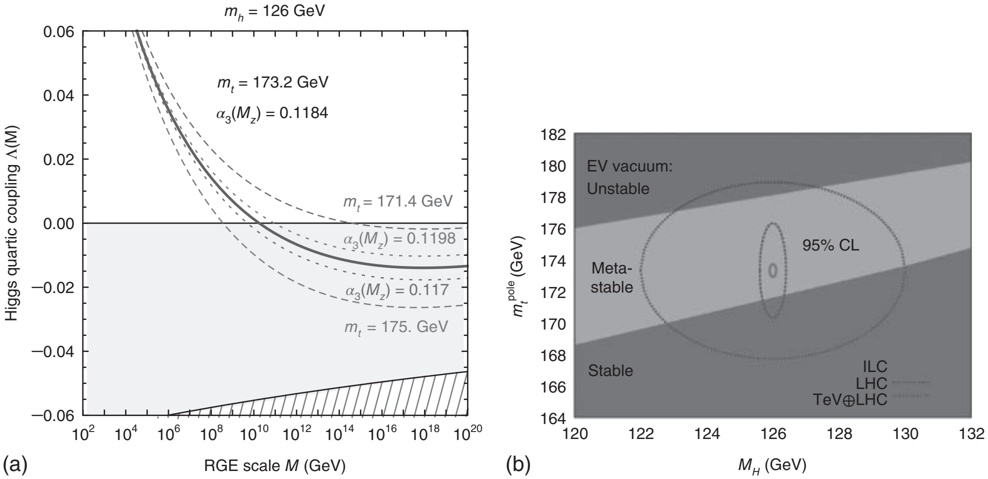


Figure 1.6 (a, b) Renormalization group evolution (RGE) of the Higgs self-coupling for $m_h = 126$ GeV for the central value of m_t and α_s, as well as for $\pm 2\sigma$ variations of m_t (dashed lines) and α_s (dotted lines). For negative values of λ, the lifetime of the SM vacuum due to quantum tunneling at zero temperature is longer than the age of the Universe as long as λ remains above the region shaded in red, which takes into account the finite corrections to the effective bounce action renormalized at the same scale as λ. (Reproduced with permission of [8, 22].) The 2σ ellipses in the $m_H - m_{\text{top, pole}}$ plane are drawn for the three cases that one obtains from the current top quark and Higgs mass measurements at the Tevatron and LHC, and which can be expected in future measurements at the LHC and at the ILC. The colored area denotes where the SM vacuum is absolutely stable, metastable, and unstable up to the Planck scale. (Reproduced with permission of [16, 23].) (Please find a color version of this figure on the color plates.)

$\sim$125 GeV would give vacuum stability up to only scales between 10^9 and 10^{10} GeV, and stability up to the Planck scale would require new physics. Such new physics could be the SUSY, but other models have also been discussed [25, 26].

This is the reason why the discovery and the determination of the Higgs mass were so important. As the dynamical properties of the Higgs are directly related to the framework of the SM, it is necessary to know how it is produced. We have to be aware that the production and detection mechanisms differ depending on the mass of the Higgs. They also differ if an $e^- e^+$ collider is used. We shall review how to produce and detect the Higgs in section 1.7.

1.4
Little and Big Hierarchy Problem

As we emphasized at the beginning of this chapter, the Higgs sector is the least known territory in the SM. All we know is that the gauge symmetry is broken by some VEV referred to as the *Higgs mechanism*. The Higgs Lagrangian in the

SM was chosen for its simplicity, satisfying the minimum requirements for the spontaneous symmetry breaking. We have no idea about the dynamic properties of the Higgs. We do not know whether the Higgs field is an elementary particle, or a composite, or a representation of some dynamical phenomenon. Because of the lack of detailed knowledge of the mechanism of the EW phase transition, the SM loses its predictability for phenomena beyond the teraelectronvolt (TeV) range. It had been expected that new physics would appear at the energy scale of $\sim O(1)$ TeV. However, from the analysis of the electroweak precision data (EWPD) obtained at LEP and Tevatron, reproducibility of the SM is so accurate that any possible deviation due to new physics has to be suppressed by making its cutoff (i.e., energy scale) larger than $\sim O(10)$ TeV. The same EWPD constrains the Higgs mass to be light, ~100 GeV with upper limit $m_h < 219$ GeV [7, 27], as we saw in Figure 1.3. Indeed, the mass value of the discovered Higgs turned out to be ≈ 125 GeV. Let us see what problem arises by "the too good SM."

In the SM, the mass correction to a scalar particle is quadratically divergent because, unlike gauge particles or chiral fermions, there is no known symmetry to suppress the divergence.

Now, if one tries to make corrections to the Higgs mass assuming the validity of the SM up to the energy scale ~10 TeV, it gives rise to an unacceptably large Higgs mass value.

This can be seen as follows. There are three types of radiative corrections to the Higgs mass that arose from the diagrams in Figure 1.7a–c.

Each of them gives a correction to the Higgs mass [28]:

$$\text{top loop} \qquad -\frac{3}{8\pi^2}y_t^2\Lambda^2 \qquad \sim -(2\,\text{TeV})^2 \tag{1.39a}$$

$$\text{gauge loop} \qquad +\frac{1}{16\pi^2}g^2\Lambda^2 \qquad \sim +(0.7\,\text{TeV})^2 \tag{1.39b}$$

$$\text{Higgs loop} \qquad +\frac{1}{16\pi^2}\lambda^2\Lambda^2 \qquad \sim +(0.5\,\text{TeV})^2 \tag{1.39c}$$

The numbers in the third column indicate the values necessary to keep the Higgs mass within the phenomenological limit. If we insert $\Lambda = 10$ TeV in the above expressions, they give

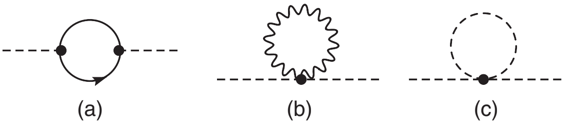


Figure 1.7 Radiative corrections due to the Standard Model. (a) Yukawa coupling with the top quark. (b) Gauge boson loop. (c) Higgs quartic self-interaction.

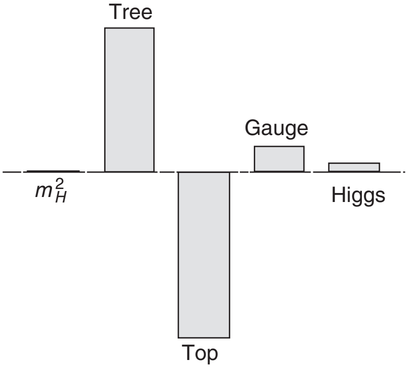


Figure 1.8 Fine-tuning. Three different and large corrections end up with a small Higgs mass.

$$\therefore \quad m_h^2 \simeq m_{\text{tree}}^2 - (100 - 10 - 5) \times (200\,\text{GeV})^2 \tag{1.40}$$

that is, the three large corrections have to conspire to give the small Higgs mass value of ~100 GeV. It is a fine-tuning of about 1 part in 100 (see Figure 1.8).

If one avoids the fine-tuning, then one has to limit the energy scale of the new physics below 1–2 TeV. This is referred to as the *little hierarchy problem*. Obviously, the Higgs mass is protected by some kind of symmetry.

The big hierarchy problem appears when one goes to GUTs, that is, if one tries to unify the strong interactions with the EW interactions. As will be discussed in Chapter 3, the grand unification occurs at the energy scale $\sim 10^{16}$ GeV, which is referred to as the *GUT scale*. Extra gauge particles, denoted as X, Y, and Higgs particles denoted as Φ, would also appear.

In the GUTs, the first phase transition occurs at the GUT scale $E_{\text{GUT}} \sim 10^{16}$ GeV and separation of the strong and EW interactions ensue. The second phase transition occurs at the EW scale $E_{\text{EW}} \sim 1/\sqrt{\sqrt{2}G_F} = 246$ GeV, which causes the EW symmetry breakdown of the SM. The extra gauge and Higgs particles acquire masses at the first phase transition.

As energy scale of radiative corrections are generally of the same order as the mass scale of the participating fields, if the Higgs is to fulfill its role of the EW mass generator as an elementary particle, its mass value should be of the order of the EW energy scale. This means that the GUT radiative corrections due to extra gauge bosons X, Y, and heavy Higgs have also to be suppressed to the same scale. This is a fine-tuning to the precision of $(m_{\text{EW}}/m_{\text{GUT}})^2 \sim 10^{28}$. This has to be done at every order of the perturbation expansion that is considered unnatural. The problem is referred to as the *big hierarchy problem* to distinguish it from the little hierarchy problem we have described above.

There are three main approaches to this problem. The first is to introduce a new symmetry, that is, the SUSY, and to try to cancel the diverging radiative corrections order by order by introducing new particles. This is a perturbative approach, assuming the Higgs to be elementary. The second is to assume that a new strong interaction that breaks the EW symmetry dynamically provides a form factor (i.e., unbinding), with the binding energy playing the role of the cutoff. This is a nonperturbative approach assuming the Higgs to be a composite particle. The third approach is to introduce extra dimensions, a possibility that will be treated in detail in Chapter 6. The second approach generally leads to a heavy Higgs. Therefore, the preferred choice is the first approach. Still, there are models that allow light Higgs (little Higgs model) in the dynamical symmetry-breaking framework.

1.5
Higgs in the Supersymmetry

The SUSY connects fermions and bosons. It introduces a new partner to every particle in the SM. The partners have spins differing by 1/2 and the same coupling strength such that their additional contributions to the mass corrections cancel those due to the SM partners. The SUSY approach is most attractive because it has nice features in addition to solving the little as well as the big hierarchy problem. By gauging the SUSY, it can produce gravity. Thus, it has the potential to unify all the four forces. Moreover, it is capable of making many definitive predictions that can be tested experimentally. Phenomenologically, it is the most comprehensively studied branch of possible new physics. Details of the SUSY is discussed in Chapters 4 and 5. Here we limit our discussions to its aspect that is specifically related to the Higgs properties.

1.5.1
Two Higgs Doublets

Properties of the Higgs particle in the SM are determined by the Higgs potential. Its form was chosen for its simplicity with minimum requirements, that is, positive quartic coupling $\lambda > 0$, to stabilize the vacuum, and negative quadratic coupling $\mu^2 = -\lambda v^2 < 0$ to induce spontaneous symmetry breaking. So far, there are no observables that contradict this assumption. Besides, the minimum model has been an important guiding principle in exploring the origin and characteristics of spontaneous symmetry breaking theoretically as well as in planning experiments to discover the Higgs particles. However, with EWPD and GUTs in mind, it is important to consider more generic models allowed by phenomenology. There are two important observational constraints in extending the SM.

(1) $\rho = \dfrac{G_N}{G_F} = \dfrac{m_W^2}{m_Z^2 \cos^2\theta_W} \approx 1.$

(2) No FCNC (flavor-changing neutral current)

The first is the neutral to charged coupling constant ratio which stays close to unity despite many radiative corrections. A custodial $SU(2)$ symmetry (see Section 3.3.5 of [2]) is believed to be active to protect it. For the second, it is guaranteed by the unitarity of the CKM matrix (Glashow–Iliopoulos–Maiani, GIM, mechanism) in the SM, but if one wants a new physics, an extended GIM mechanism is required.

At the tree level, condition (1) is satisfied automatically if the additional Higgs is a member of doublets (see Eq.(5.12c) of [2]). Therefore, the simplest extension is to require an extra Higgs doublet. The SUSY also requires a minimum of two Higgs doublets. When two Higgs doublets are assumed, there are several ways to satisfy condition (2). The most conventional assumption is to require that one Higgs doublet couples only to up-type quarks (u,c,t) with $Q = +2/3$ and the other only to down-type quarks (d,s,b) with $Q = -1/3$. Let Φ_1, Φ_2 denote two complex $Y = 1$, $SU(2)$ doublet scalar fields. The most general $SU(2)$ gauge-invariant scalar potential can be written down as [5]

$$\begin{aligned}V =& m_{11}^2\Phi_1^\dagger\Phi_1 + m_{22}^2\Phi_2^\dagger\Phi_2 - \{m_{12}^2\Phi_1^\dagger\Phi_2 + h.c.\}\\ &+ \frac{1}{2}\lambda_1(\Phi_1^\dagger\Phi_1)^2 + \frac{1}{2}\lambda_2(\Phi_2^\dagger\Phi_2)^2 + \lambda_3(\Phi_1^\dagger\Phi_1)(\Phi_2^\dagger\Phi_2) + \lambda_4(\Phi_1^\dagger\Phi_2)(\Phi_2^\dagger\Phi_1)\\ &+ \left\{\frac{1}{2}\lambda_5(\Phi_1^\dagger\Phi_2)^2 + \left[\lambda_6(\Phi_1^\dagger\Phi_1) + \lambda_7(\Phi_2^\dagger\Phi_2)\right]\Phi_1^\dagger\Phi_2 + h.c.\right\}\end{aligned} \tag{1.41}$$

All the coefficients are real if charge parity (CP) invariance is assumed. We simplify the above potential by imposing the SUSY. In order not to make the arguments too complicated, we adopt the MSSM (minimum supersymmetric extension of the Standard Model) assumption here.

In MSSM, one of the doublets, which we denote as H_1, has the same quantum number ($Y = -1$) as the charge conjugate of Φ_1 and couples to down-type quarks. The other, denoted as H_2, has the quantum number of Φ_2 ($Y = +1$) and couples to up-type quarks. As to what the MSSM is, we defer the discussion to Chapter 5. Here we treat the MSSM Higgs potential as given $\Big($see Eq. (5.9) [29]$\Big)$. The MSSM relates the coefficients of the quartic potential with the gauge coupling, simplifying the potential to

$$\begin{aligned}V_H = &\mu_1^2 H_1^\dagger H_1 + \mu_2^2 H_2^\dagger H_2 - \mu_3^2(\varepsilon_{ij}H_1^i H_2^j + h.c.)\\ &+ \frac{g_W^2 + g_B^2}{8}(H_1^\dagger H_1 - H_2^\dagger H_2)^2 + \frac{g_W^2}{2}|H_1^\dagger H_2|^2\end{aligned} \tag{1.42}$$

where $\varepsilon_{12} = -\varepsilon_{21} = 1$, $\epsilon_{11} = \epsilon_{22} = 0$ and H_1^i, H_2^j are components of the Higgs doublet H_1 and H_2. g_W, g_B are the gauge coupling constants of $SU(2)_L$ and $U(1)_Y$, respectively. As H_1^c and H_2 have the same quantum number, we define two complex neutral and two charged scalar fields by

$$H_1 = \begin{bmatrix} H_1^1 \\ H_1^2 \end{bmatrix} = \Phi_1^c = \begin{bmatrix} \phi_1^{0\dagger} \\ -\phi_1^- \end{bmatrix} \quad \text{8)} \tag{1.43a}$$

$$H_2 = \begin{bmatrix} H_2^1 \\ H_2^2 \end{bmatrix} = \Phi_2 = \begin{bmatrix} \phi_2^+ \\ \phi_2^0 \end{bmatrix} \tag{1.43b}$$

Then, the potential Eq. (1.42) is rewritten as follows:

$$\begin{aligned} V_H &= \mu_1^2(|\phi_1^0|^2 + |\phi_1^-|^2) + \mu_2^2(|\phi_2^0|^2 + |\phi_2^+|^2) - \mu_3^2(\phi_1^{0\dagger}\phi_2^0 + \phi_1^-\phi_2^+ + h.c.) \\ &+ \frac{g_W^2 + g_B^2}{8}\left(|\phi_1^0|^2 + |\phi_1^-|^2 - |\phi_2^0|^2 - |\phi_2^+|^2\right)^2 + \frac{g_W^2}{2}\left|\phi_1^{0\dagger}\phi_2^+ - \phi_1^+\phi_2^0\right|^2 \end{aligned} \tag{1.46}$$

To break the symmetry spontaneously, we choose the VEVs to satisfy $<\phi_1^+>=<\phi_2^->= 0$ so that the charge conservation is respected. Then, to obtain the final VEV, it is enough to consider only the neutral fields. We assume that the minimum of the potential is at

$$<H_1^c>= \begin{bmatrix} 0 \\ \frac{v_1}{\sqrt{2}} \end{bmatrix}, \qquad <H_2>= \begin{bmatrix} 0 \\ \frac{v_2}{\sqrt{2}} \end{bmatrix} \tag{1.47}$$

There are two important conditions for the potential:

(1) For the vacuum stability, the potential has to be bound from below. For $|\phi_1^0| \neq |\phi_2^0|$, it is automatic, but for $|\phi_1^0| = |\phi_2^0|$, one needs a condition $\mu_1^2 + \mu_2^2 > 2\mu_3^2$.
(2) To obtain symmetry breaking, the coefficient of the quadratic term must be negative, which amounts to $\mu_3^4 > \mu_1^2\mu_2^2$.

The minimum of the potential can be obtained by substituting $\phi_1^0 = v_1/\sqrt{2}$, $\phi_2^0 = v_2/\sqrt{2}$ in the potential and requiring $\partial V_H/\partial v_1 = \partial V_H/\partial v_2 = 0$. We obtain two

---

8) This way of arranging the Higgs doublets may strike odd for those who are accustomed to the four-component Dirac spinor representation. Indeed, in [1] and [2] of this book (and also in Equation (1.7)) , the Yukawa interaction to give mass to a quark doublet is arranged as follows:

$$\begin{aligned} -\mathscr{L}_{\text{Yukawa}} &= y_d(\overline{\Psi}_L\Phi)d_R + y_u(\overline{\Psi}_L\Phi_c)u_R + h.c. \\ \Psi_L &= \begin{bmatrix} u_L \\ d_L \end{bmatrix}, \quad \Phi = \begin{bmatrix} \phi^+ \\ \phi^0 \end{bmatrix}, \Phi_c = \begin{bmatrix} \phi^{0\dagger} \\ -\phi^- \end{bmatrix} \end{aligned} \tag{1.44}$$

The reason why the Higgs fields are expressed like Eq. (1.43) is because, in the supersymmetric formalism, two-component left-handed Weyl spinors are the preferred bases to express fermion fields. Both quark and Higgs doublets are in the left-handed representation. The SU(2) invariant made of two doublets is expressed as (see Appendix A of [1])

$$\begin{aligned} &- y_u(u_R)^c\varepsilon_{ij}q_L^iH_2^j + y_d(d_R)^c\varepsilon_{ij}q_L^iH_1^j, \\ &(q_L^1, q_L^2) = (u_L, d_L), \ \varepsilon_{12} = -\varepsilon_{21} = -1 \end{aligned} \tag{1.45}$$

Thus, the position of the neutral component in the Higgs doublets should be reversed to produce the correct mass terms when they get the vacuum expectation value.

relations as

$$\begin{aligned}\mu_1^2 &= \mu_3^2\frac{v_2}{v_1} - \frac{1}{8}(g_W^2 + g_B^2)(v_1^2 - v_2^2) = \mu_3^2 \tan\beta - \frac{m_Z^2}{2}\cos 2\beta \\ \mu_2^2 &= \mu_3^2\frac{v_1}{v_2} + \frac{1}{8}(g_W^2 + g_B^2)(v_1^2 - v_2^2) = \mu_3^2 \cot\beta + \frac{m_Z^2}{2}\cos 2\beta\end{aligned} \tag{1.48}$$

where the important angle β is defined by

$$\tan\beta \equiv \frac{v_2}{v_1} \tag{1.49}$$

The VEV of the SM, $v^2 \equiv v_1^2 + v_2^2$, is fixed by the relation $m_W = g_W v/2$, which leads to the second equality in Eqs. (1.48). The Z mass-squared, in turn, is given by

$$m_Z^2 = \frac{\mu_1^2 - \mu_2^2 \tan^2\beta}{\tan^2\beta - 1} \tag{1.50}$$

One sees that the two equations in Eq. (1.48) automatically satisfy the vacuum stability and the symmetry-breaking conditions.

1.5.2
Coupling Strengths of MSSM Higgs

As the SUSY relates the gauge sector to the Higgs sector, the coupling of the Higgs is uniquely determined once the SUSY parameters ($\tan\beta, \tan\alpha$) are given. α is the mixing angle of neutral Higgs fields and is given by Eq. (1.63). As it is lengthy to write down all the interactions, we list a few in Tables 1.1–1.4 [4, 30] to illustrate the constraints imposed by the SUSY.

Here, $\varphi_{\rm SM}$ denotes the Higgs in the SM.

The coupling of the charged Higgs to the fermion is given by

$$\mathscr{L}_{\rm INT} = \frac{g_W}{2\sqrt{2}m_W}\bar{t}\left[m_t \cot\beta(1+\gamma^5) + m_b \tan\beta(1-\gamma^5)\right] b H^+ + h.c. \tag{1.51}$$

This equation is written in terms of (t,b), but couplings to other fermions can be written down similarly. The couplings being specified, all the remaining parameters are the masses of the particle. Therefore, the production cross section, and hence the detection method, can be determined once the mass value is specified.

Table 1.1 Coupling strength of the neutral Higgs to the fermion.

$$\mathscr{L}_{\rm INT} \sim -\frac{g_W}{2}\frac{m_f}{m_W}\bar{f}O_i f\phi_i$$

ϕ_i		h^0	H^0	A	$\varphi_{\rm SM}$
O_i	$f = t$	$\frac{\cos\alpha}{\sin\beta}$	$\frac{\sin\alpha}{\sin\beta}$	$-i\gamma^5\cot\beta$	1
	$f = b$	$-\frac{\sin\alpha}{\cos\beta}$	$\frac{\cos\alpha}{\cos\beta}$	$-i\gamma^5\tan\beta$	1

Table 1.2 Coupling strength of the neutral Higgs to the vector boson.

$$\mathscr{L}_{\text{INT}} \sim ig_V m_V O_i g_{\mu\nu} V_\mu V^\nu \phi_i$$

$$(g_W = e/\sin\theta_W,\ g_Z = e/\sin\theta_W \cos\theta_W)$$

ϕ_i	h^0	H^0	A	φ_{SM}
O_i	$-\sin(\alpha-\beta)$	$\cos(\alpha-\beta)$	0	1

Table 1.3 Coupling strength of the neutral CP-odd Higgs to the vector boson.

$$\mathscr{L}_{\text{INT}} \sim V\phi(p)\phi(p') \sim \tfrac{g_Z}{2} O_i (p-p')^\mu Z_\mu A^0 \phi_i$$

ϕ_i	h^0	H^0	φ_{SM}
O_i	$\cos(\alpha-\beta)$	$-\sin(\alpha-\beta)$	0

Table 1.4 Coupling strength of the charged Higgs to the vector boson.

$$\mathscr{L}_{INT} \sim V\phi(p)\phi(p') \sim -i\tfrac{g_W}{2} O_i (p-p')^\mu W^\pm_\mu H^\mp \phi_i$$

ϕ_i	h^0	H^0	φ_{SM}
O_i	$\cos(\alpha-\beta)$	$\sin(\alpha-\beta)$	0

We give a partial list of allowed and forbidden couplings at the tree level. The radiative corrections modify them, but they are omitted.

$$\begin{array}{ll} \bigcirc: & [VVH^0,\ VVh^0], \quad [ZA^0H^0,\ ZA^0h^0], \quad ZH^+H^-, \quad [W^\pm H^\mp H^0,\ W^\pm H^\mp h^0] \\ & [ZZH^0,\ ZZh^0], \cdots \\ \times: & VVA^0,\ ZW^\pm H^\mp,\ ZH^0h^0,\ ZH^0H^0,\ Zh^0h^0,\ ZA^0A^0, \cdots \end{array} \tag{1.52}$$

The coupling of the pair in the $[\cdot\cdot, \cdot\cdot]$ is like $\sim [\sin x, \cos x]$ and complementary in the sense that, if one is smaller, the other is larger. The complementarity is related to the unitarity. In a theory where the symmetry is spontaneously broken, processes containing the Higgs have a role to compensate a class of diverging integrals (see discussions in Chapter 1 of [2]). For example, if a process in which both H^0, h^0 are exchanged contributes to compensation of diverging WW integrals, the combined effect of the H^0, h^0 has to be the same as the SM's Higgs which constrains the relation between the two. That is,

$$g^2(h^0WW) + g^2(H^0WW) = g_W^2 \tag{1.53}$$

Referring to Table 1.2, one sees that the above equation is satisfied. Furthermore, we also have

$$g^2(h^0ZZ) + g^2(H^0ZZ) = g_Z^2 \tag{1.54a}$$

$$g^2(h^0A^0Z) + g^2(H^0A^0Z) = g_Z^2 \tag{1.54b}$$

These considerations mean that one of A^0, h^0, H^0 has a detection probability similar to that of the SM Higgs.

1.5.3
Mass Spectrum of MSSM Higgs

Spontaneous breakdown of the symmetry induces mixing and changes the mass eigenvalues. Out of eight scalar fields, three are taken up by the gauge bosons to give them mass, and five appear as physical particles. Three of them are neutral and the rest two are charged. If the CP invariance is assumed, the real and imaginary parts of the scalar fields do not mix and the charged Higgs's are also separated. Let us derive the mass of the CP-odd neutral scalar first. By extracting terms quadratic in $\text{Im}\,\phi_1^0$ and $\text{Im}\,\phi_2^0$, we obtain the following mass matrix:

$$M_A^2 = \mu_3^2 \times (\text{Im}\,\phi_1^0, \text{Im}\,\phi_2^0) \begin{bmatrix} \frac{v_2}{v_1} & -1 \\ -1 & \frac{v_1}{v_2} \end{bmatrix} \begin{pmatrix} \text{Im}\,\phi_1^0 \\ \text{Im}\,\phi_2^0 \end{pmatrix} \tag{1.55}$$

As the determinant vanishes, one eigenvalue is 0 corresponding to the would-be-Goldstone boson which was eaten by the Z boson.[9] The other corresponds to a CP-odd neutral scalar field, which we denote as A. By diagonalizing the mass matrix, we obtain masses and expressions for the field A as well as the would-be-Goldstone field G^0.

$$m_A^2 = \mu_3^2(\tan\beta + \cot\beta) = \frac{2\mu_3^2}{\sin 2\beta} \tag{1.56}$$

$$A = \sqrt{2}\left(-\text{Im}\phi_1^0 \sin\beta + \text{Im}\phi_2^0 \cos\beta\right), \tag{1.57a}$$

$$G^0 = \sqrt{2}(\text{Im}\phi_1^0 \cos\beta + \text{Im}\phi_2^0 \sin\beta) \tag{1.57b}$$

The mass matrix of the charged Higgs can be obtained similarly.

$$\begin{aligned} M_{H^\pm}^2 &= \left(\frac{\mu_3^2}{v_1 v_2} + \frac{g_W^2}{4}\right) \times (\phi_1^+, \phi_2^+) \begin{bmatrix} v_2^2 & -v_1 v_2 \\ -v_1 v_2 & v_1^2 \end{bmatrix} \begin{pmatrix} \phi_1^- \\ \phi_2^- \end{pmatrix} \\ &= (m_A^2 + m_W^2) \times (\phi_1^+, \phi_2^+) \begin{bmatrix} \sin^2\beta & -\sin\beta\cos\beta \\ -\sin\beta\cos\beta & \cos^2\beta \end{bmatrix} \begin{pmatrix} \phi_1^- \\ \phi_2^- \end{pmatrix} \end{aligned} \tag{1.58}$$

where we used Eq. (1.49) and (1.56) in going to the second line. Again, one of the masses vanishes. The other mass has a finite eigenvalue and is physical. They are related by

$$m_{H^\pm}^2 = m_A^2 + m_W^2 \tag{1.59}$$

$$H^\pm = -\phi_1^\pm \sin\beta + \phi_2^\pm \cos\beta, \qquad G^\pm = \phi_1^\pm \cos\beta + \phi_2^\pm \sin\beta \tag{1.60}$$

---

[9] The Goldstone bosons that are absorbed by gauge particles are referred to as the *would-be-Goldstone bosons*, and physical Goldstone bosons that acquire mass through additional external force are referred to as the *pseudo Nambu-Goldstone bosons* (pNGBs) to distinguish them from the zero-mass Goldstone bosons.

In order to obtain the mass matrix of the two CP-even neutral Higgs particles, we set $\phi_1^0 = v_1/\sqrt{2}$, $\phi_2^0 = v_2/\sqrt{2}$, $\phi_1^\pm = \phi_2^\pm = 0$ and calculate

$$\begin{aligned} M_{\text{CP+}}^2 &= \frac{1}{2}\left.\frac{\partial^2 V_H}{\partial v_i \partial v_j}\right|_{v_i, v_j = 0} = \begin{bmatrix} \mu_3^2 \tan\beta + m_Z^2 \cos^2\beta & -\mu_3^2 - m_Z^2 \sin\beta\cos\beta \\ -\mu_3^2 - m_Z^2 \sin\beta\cos\beta & \mu_3^2 \cot\beta + m_Z^2 \sin^2\beta \end{bmatrix} \\ &= \begin{bmatrix} m_A^2 \sin^2\beta + m_Z^2 \cos^2\beta & -\frac{1}{2}(m_A^2 + m_Z^2)\sin 2\beta \\ -\frac{1}{2}(m_A^2 + m_Z^2)\sin 2\beta & m_A^2 \cos^2\beta + m_Z^2 \sin^2\beta \end{bmatrix} \end{aligned} \tag{1.61}$$

Diagonalizing the mass matrix, we obtain

$$m_{H,h}^2 = \frac{1}{2}\left[m_A^2 + m_Z^2 \pm \sqrt{(m_A^2 + m_Z^2)^2 - 4m_Z^2 m_A^2 \cos^2 2\beta}\right] \tag{1.62}$$

It is customary to denote the heavier one of the two mass eigenstates as H^0 and the lighter one as h^0. They are given by

$$H^0 = \left(\sqrt{2}\text{Re}\phi_1^0 - v_1\right)\cos\alpha + \left(\sqrt{2}\text{Re}\phi_2^0 - v_2\right)\sin\alpha \tag{1.63a}$$

$$h^0 = -\left(\sqrt{2}\text{Re}\phi_1^0 - v_1\right)\sin\alpha + \left(\sqrt{2}\text{Re}\phi_2^0 - v_2\right)\cos\alpha \tag{1.63b}$$

$$\tan 2\alpha = \frac{m_A^2 + m_Z^2}{m_A^2 - m_Z^2}\tan 2\beta, \qquad -\frac{\pi}{2} \le \alpha \le 0 \tag{1.63c}$$

From the above equations, we have the relations

$$\begin{aligned} &m_{H^\pm}^2 > m_W^2 \\ &m_{H^0} > \max(m_A, m_Z) \ge \min(m_A, m_Z)\cos 2\beta \ge m_h \\ &m_h^2 + m_{H^0}^2 = m_Z^2 + m_A^2 \end{aligned} \tag{1.64}$$

Thus, the supersymmetric structure of the theory has imposed very strong conditions on the Higgs spectrum. Out of six parameters that describe the MSSM Higgs sector $(m_h, m_H, m_A, m_{H^\pm}, \beta, \alpha)$, only two parameters that can be taken as $\tan\beta$ and m_A are free parameters at the tree level.

The relations Eqs. (1.64) mean that at least one of the three neutral Higgs particles is lighter than Z. Therefore, if one can prove the existence of a light Higgs h^0 with mass smaller than Z, the possibility of the SUSY, or at least its minimum version MSSM, is very large. Notice, however, that this story is valid only for the tree approximation. Inclusion of radiative corrections modifies the mass value, especially those including the massive top and its spin 0 partner "stop" quark. The correction to the light Higgs mass is given by [31–34] [30, 35]

$$\begin{aligned} m_h^2 &\to m_h^2 + \delta m_h^2 \\ &\simeq m_Z^2 + \frac{3}{2\pi^2}\left(\frac{m_t^4}{v^2}\right)\left[\ln\left(\frac{M_S^2}{m_t^2}\right) + \frac{X_t^2}{M_S^2}\left(1 - \frac{X_t^2}{12M_S^2}\right)\right]^{10)} \end{aligned} \tag{1.65}$$

where $X_t = A_t - \mu \cot\beta$ is the mixing parameter in the stop sector [see Eq. (5.35)]. We have characterized the scale of "stop" ($\tilde{t}_1, \tilde{t}_2$) masses with $M_S \equiv (m_{\tilde{t}_1} m_{\tilde{t}_2})^{1/2}$.

Figure 1.9 shows the values of corrected m_h as well as m_{H^0}, $m_{H^\pm}$ as a function of m_A for two values of $\tan\beta = 3, 30$ with the maximum mixing scenario. m_h almost saturates and is insensitive to $\tan\beta$ or m_A whether they take values larger than those given in Figure 1.9. Thus one sees that, even after the radiative corrections, the mass of the light Higgs remains relatively small, $m_h \lesssim 120 \sim 140$ GeV. The existence of the light Higgs is a solid prediction of the SUSY. The mass value ($m_h = 125.7 \pm 0.4$ GeV) determined by LHC is near the upper limit of the SUSY predictions. It means that parameters of the MSSM to reproduce the observed value are near the boundaries of allowed regions. The stop mass, in particular, takes a large value ($\gtrsim$ a few tetraelectron volts) in the conventional scenario of the MSSM [36].

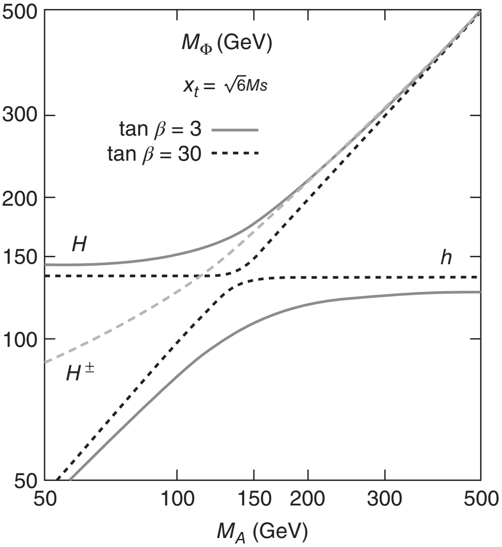


Figure 1.9 Mass of the MSSM Higgs bosons as functions of m_A for two values of $\tan\beta = 3, 30$ with maximal mixing scenario with the stop ($\tilde{t}$) mass $m_{\tilde{t}} = 2$ TeV and all other SUSY parameters set to 1 TeV. (Reproduced with permission of [30].)

---

[10] There are two "stop" mass eigenstates $\tilde{t}_1, \tilde{t}_2$ that are obtained by the mixing of $\tilde{t}_L$, $\tilde{t}_R$, which are super partners of the top (t_L, t_R). M_S is defined as the geometric mean of the two stop masses $\left(M_S = \sqrt{\{m(\tilde{t}_1)m(\tilde{t}_2)\}}\right)$. The mixing is induced by the Yukawa coupling of the Higgs to both particles.

In the SM, the Higgs mass is basically a free parameter. In the MSSM model, however, the light Higgs (h) is bounded from above with its mass given by Eq. (1.65). This is obtained in the so-called decoupling regime where the value of m_A is set high, pushing masses of other Higgses (H^0, $H^{\pm}$) high also. This is a scenario to maximize m_h in the framework of SUSY. An analysis of the Higgs mass guided by the naturalness condition (i.e., no excessive fine tuning) is shown in Figure 1.10a. One sees that a large mixing (large $A_t - \mu \cot\beta$, see Eqs. (1.65), (5.35)) as well as a large stop mass is required to realize the observed value of the Higgs. As the lighter stop is considered as the lightest of all squarks, it means that other squark masses are at least higher than ~ 700GeV, consistent with direct search results. The difference of the two curves in the figure (Suspect and FeynHiggs) may be considered as the theoretical uncertainty. Validity of various SUSY models was also examined and is shown in Figure 1.10b. The observed Higgs mass value excludes the GMSB (gauge mediated symmetry breaking) and AMSB (anomaly mediated symmetry breaking) in their simplest version. They will be discussed in Section 5.4 and 5.5. For no or small mixing, a much higher value of M_S is required. The limit of $M_S < 3$ TeV may be set from naturalness consideration. If one removes the constraint on these models, freedom of realizing the observed mass increases, and many alternative models are being discussed.

In the high scale SUSY, for instance, the mass scale of all the SUSY particles are set high, while in the split SUSY [40–42], only the scalars (squarks and sleptons) are pushed beyond the LHC reach ($\sim 10^9$ GeV). Figure 1.11 shows the predicted range of the Higgs mass [16]. However, fermions, that is, gauginos and higgsinos, remain at the low energy scale. One conspicuous feature of the split SUSY is a long-lived

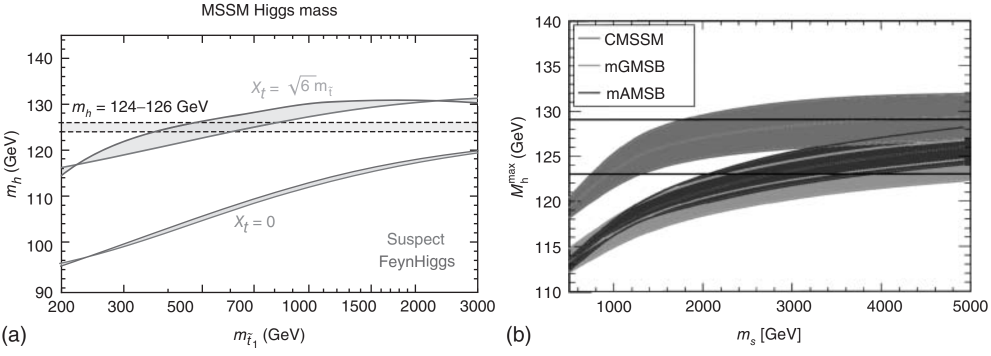


Figure 1.10 (a) Higgs mass as a function of the scalar top mass. Other SUSY parameters were fixed at values guided by naturalness conditions. Large mixing and stop mass are required. (Reproduced with permission of [37]. (b) The Higgs mass shown as a function of $M_S = (m_{\tilde{t}_1} m_{\tilde{t}_2})^{1/2}$ for the various constrained MSSM models. Note that, for reasonable values of $M_S \lesssim 3000$ GeV, simple versions of GMSB, AMSB are ruled out. CMSSM (constrained minimum supersymmetric extension of the Standard Model discussed in Section 5.3) survives. (Reproduced with permission of [38, 39]. (Please find a color version of this figure on the color plates.)

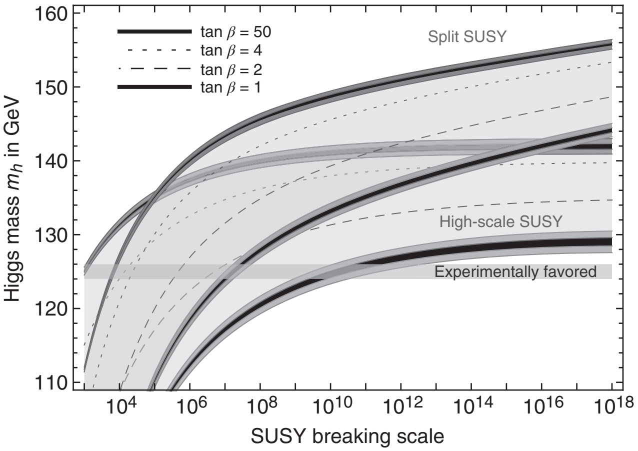


Figure 1.11 NNLO prediction for the Higgs mass m_h in high-scale supersymmetry (blue, lower) and split supersymmetry (red, upper) for $\tan\beta = \{1, 2, 4, 50\}$. The thickness of the lower boundary at $\tan\beta = 1$ and of the upper boundary at $\tan\beta = 50$ shows the uncertainty due to the present 1σ error on α_s (black band) and on the top mass (larger colored band). (Reproduced with permission of [16].) (Please find a color version of this figure on the color plates.)

gaugino, which could have lifetime as long as 100 s [43, 44]. This is because the gluino's decay has to go through the heavy squark loops. However, the split SUSY retains basic virtues of the SUSY, that is, gauge unification (see Section 3.6), the dark matter candidate and possibly the light SM-like Higgs boson [45].

In summary, the mass of the Higgs is still within the allowed limit of the MSSM, but many parameters are pushed almost to their extreme limits, thereby excluding many of the more constrained models.

1.6
Is the Higgs Elementary?

So far, we treated the Higgs as an elementary particle. Notice, however, that the Higgs mechanism was constructed using superconductivity as a model. In superconductivity, spontaneous symmetry breaking is induced by the Cooper pair, which is a composite of two electrons. In QCD, chiral symmetry breaking is induced by the condensate of the quark–antiquark pair. It is quite logical to think that the Higgs may also be a composite. In this case, one considers a new, strong interaction which works among new particles and regards the Higgs as a bound state of some particle pairs.[11] Representative models are the technicolor (TC) model and, more recently, the little Higgs model.

---

[11] One does not necessarily introduce new particles or new interactions. For instance, the top condensate model considers the Higgs as a bound state of top and anti-top quarks [46–48].

1.6.1
Technicolor Model

A representative theory of the strong dynamical symmetry breaking is the TC model [49–52]. Just like the pion is the NGB, which emerges from the chiral symmetry breaking $\left(SU(2)_L \times SU(2)_R \to SU(2)_V \right)$ as a result of the $q\bar{q}$ condensation in QCD, the Higgs can also be considered as the NGBs of the new gauge symmetry $SU(N_{\text{TC}})$ with new fermions (techni-fermions)

$$\Psi_L = \begin{bmatrix} U \\ D \end{bmatrix}_L, \quad U_R, \quad D_R \tag{1.66}$$

constituting the fundamental representation. If $N_{\text{TC}} = 3$, one may consider TC as the scaled-up version of QCD = color $SU(3)$. For simplicity, however, we consider $N_{\text{TC}} = 1$, with only two flavors (U, D). Generalization to $N_{\text{TC}} > 1$ can be done easily. The fermion kinetic energy terms for this theory are

$$\mathscr{L}_{\text{kin}} = \overline{\Psi}_L i D_\mu \gamma^\mu \Psi_L + \overline{U}_R i D_\mu \gamma^\mu U_R + \overline{D}_R i D_\mu \gamma^\mu D_R \tag{1.67}$$

and like QCD, they have a chiral $SU(2)_{\text{L}} \times SU(2)_{\text{R}}$ symmetry in the $m_U = m_D = 0$ limit. The gauge boson is referred to as the *techni-gluon*, and the interaction by exchange of the techni-gluon induces the formation of a condensate

$$<\overline{U}_L U_R>=<\overline{D}_L D_R>\simeq F_{\text{TC}}^3 \tag{1.68}$$

which dynamically breaks the gauge symmetry $SU(2)_L \times SU(2)_R \to SU(2)_V$. Just like the QCD, the NGBs appear as the techni-pions $(\pi_{\text{TC}}^+, \pi_{\text{TC}}^0, \pi_{\text{TC}}^-) = [\overline{D}U, (\overline{U}U - \overline{D}D)/\sqrt{2}, \overline{U}D]$, which are absorbed by the gauge bosons of the EW interaction.[12] When there are N_D techni-fermion doublets, the constant F_{TC} is modified to [52]

$$F_{\text{TC}} \simeq \frac{v_0}{\sqrt{N_D}}, \qquad v_0 = 246\,\text{GeV} \tag{1.69}$$

for $N_D = 1$ $F_{\text{TC}} = v_0$. In QCD, the VEV of the quark condensate is related to the pion decay constant F_π by the relation

$$\begin{aligned} <\bar{u}_L u_R> &\simeq<\bar{d}_L d_R>\simeq F_\pi^3 \\ F_\pi &= 96\ \text{MeV} \end{aligned} \tag{1.70}$$

Therefore, TC is a scaled-up QCD by a factor of

$$\frac{F_{\text{TC}}}{F_\pi} = \frac{246\,\text{GeV}}{96\,\text{MeV}} \approx 2500 \tag{1.71}$$

If the chiral symmetry is larger than $SU(2)_L \times SU(2)_R$, it will contain extra NGBs which are not eaten by the gauge particles. They will acquire mass (referred to as *pseudo-Nambu-Goldstone boson, pNGB*) due to the nonzero techni-fermion mass, just like $SU(3)_{\text{flavor}}$ octet NGBs are massive in QCD. While the Higgs may not appear in TC, it is likely that the strongly interacting pNGBs form a variety of resonances ($\rho_{\text{TC}}/\omega_{\text{TC}}$, a_{TC}, etc.),[13] just like the $\pi\pi$ pair forms the vector resonance $\rho/\omega/a_1$. In such a case, a rich spectra of new particles will appear above the TeV region.

---

[12] Notice that, in the simplest model, they have spin parity 0^-. If the observed Higgs has spin parity 0^+, one has to think of a p-wave excited state to assign the NGB boson to the Higgs. This is another complication of the TC model.

Extended technicolor: The basic idea of the TC is very attractive and solves the fine-tuning problem by providing a natural cutoff (i.e., form factor of the bound states) for the high-energy part of the radiative contributions. However, realistic models have to reproduce the mass spectra of known fermions and phenomenologically established constraints, that is, suppression of FCNC. The mass generation mechanism of the ordinary fermions (denoted as f) is through Yukawa coupling in the SM, but the Higgs in TC is a composite of the technifermions. In order to generate the fermion mass, one has to have a new interaction that couples both to ordinary and TC fermions. The standard choice is the ETC (extended technicolor), a new gauge interaction at higher energy scale than the typical TC energy, which is of the order of the EW symmetry breaking. Exchange of the ETC gluon provides the fermion–fermion interaction and generates an effective mass. On the other hand, the same ETC interaction produces the FCNC interaction, and it is hard to construct a phenomenologically viable model. A remedy is the walking TC model, which assumes that the gauge coupling α_{TC} evolves slowly (i.e., it walks, not runs).

While the TC model itself is on a tightrope phenomenologically, we emphasize that the notion of the dynamical symmetry breaking is a viable one [52]. Therefore, we look for possible signals of the TC in the LHC data. Analyses are inevitably model-dependent. We show below those based on the LSTC (low-scale technicolor) model by [53, 54]. Here, the lightest techni-hadron is the techni-pion π_{TC}. The next lightest are ρ_{TC}, ω_{TC}, and $a_{1\,\text{TC}}$, which are almost degenerate.

Experimental searches: The CMS group looked for the TC particles that would appear as an excess in WZ channels [55]. The main interest was to find $\rho_T/a_{\text{TC}} \to WZ$ as well as $W' \to WZ$, where W' is a heavy W which appears in a variety of models[14]. But, here we concentrate on $\rho_{\text{TC}}/a_{\text{TC}}$ signals which are collectively called ρ_{TC}, as they cannot be distinguished in the WZ channel. Signals are $X \to 3l + \nu$.

Figure 1.12a shows their accumulated number of events as a function of the WZ invariant mass. Formally, the invariant mass of WZ cannot be uniquely determined. However, by assuming the W to have its nominal mass, the value of the neutrino longitudinal momentum is constrained to one of the two solutions of a quadratic equation. According to Monte Carlo simulations, the smaller of the two turned out to be the right solution 75% of the time and this solution was adopted for all events. The parameter $\sin\chi$ they used is given by [56]

$$\sin\chi = \frac{F_{\text{TC}}}{v_0} = \frac{F_{\text{TC}}}{246}\,\text{GeV} \approx \frac{1}{\sqrt{N_D}} \ll 1 \tag{1.72}$$

---

[13] $I^G(J^{PC})$ of π_{TC}, ρ_{TC}, ω_{TC}, $a_{1\,\text{TC}}$ are $1^-(0^{-+})$, $1^+(1^{--})$, $0^-(1^{--})$, and $1^-(1^{++})$.

[14] Conventionally, it is assumed to have identical coupling strength as the SM W, which is referred to as the *Sequential Standard Model* (SSM).

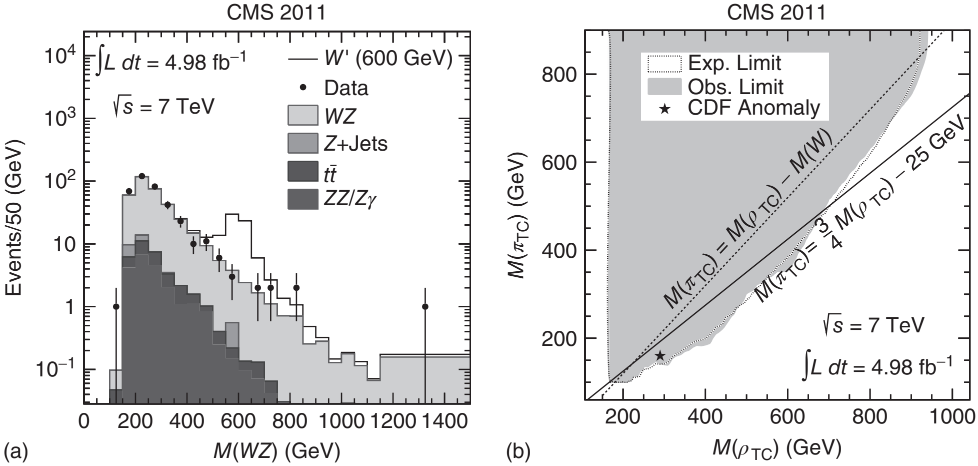


Figure 1.12 (a) Distribution of the WZ invariant mass. W' in SSM with mass point at 600 GeV is denoted as a white histogram. (b) Exclusion region at 95% CL for the LSTC (low-scale TC) model as a function of ρ_{TC} and π_{TC} masses. (Reproduced with permission of [55].

The data agreed with the SM prediction, and no signal of TC resonances was obtained. The right figure of Figure 1.12b shows the excluded region on the $M(\pi_{TC}) - M(\rho_{TC})$ plane. TC hadrons (ρ_{TC} etc.) with masses between 167 and 687 GeV have been excluded, assuming $M(\pi_{TC}) = (3/4)M(\rho_{TC}) - 25$ GeV. The region $M(\pi_{TC}) > M(\rho_{TC})$ is theoretically forbidden.

The ATLAS group, in their search for TC in dilepton decay channels, obtained similar results [57].

1.6.2
Little Higgs Model

An alternative to the TC model recently being discussed is the ‘‘little Higgs model’’ [28, 58]. Like TC, it considers the Higgs as an Nambu Goldstone Boson (NGB) produced by the spontaneously broken symmetry of a new, strong force [59,60]. But unlike TC, which introduces a new force explicitly, it focuses on global symmetry breaking, though the new interaction is implicit by its NGB assumption. It also retains the light Higgs, and is thus more realistic phenomenologically.

Denoting the global symmetry as G, it must include the EW symmetry $\left(G \supset SU(2) \times U(1)\right)$. Being the NGB, the Higgs is massless at the tree level. By treating it in the framework of the nonlinear σ model (see the boxed paragraph and also Appendix J of [2]), one can construct an effective theory of a new, strong interaction which recovers the SM in the low-energy limit.

The nonlinear σ model expresses the Nambu-Goldstone boson (NGB) as a phase field analogous to Eq. (1.12).

$$\Sigma = \sigma + i\boldsymbol{\tau}\cdot\boldsymbol{\pi} \to (v+\rho)e^{i\boldsymbol{\pi}\cdot\boldsymbol{\tau}/2v} = (v+\rho)U \tag{1.73}$$

where σ (or ρ after symmetry breakdown) is a scalar field which induces spontaneous symmetry breakdown and $\boldsymbol{\pi}$ are the NGBs associated with broken symmetry (in this case the global symmetry is $SU(2)$ chiral symmetry). In this expression, the phase transformation for $\boldsymbol{\pi}$ becomes the shift transformation, and hence the gauge invariance allows only derivative couplings for $\boldsymbol{\pi}$. Therefore, the zero mass at the tree level is automatic. The Lagrangian is expanded in powers of field derivatives but otherwise constrained only by the symmetry.

$$\mathscr{L}_{\text{eff}} = \frac{v^2}{4}\text{Tr}[\partial_\mu U\partial^\mu U^\dagger] + \frac{v^2}{8m_\rho^2}\left(\text{Tr}[\partial_\mu U\partial^\mu U^\dagger]\right)^2 + \cdots \tag{1.74}$$

As the field derivatives are momenta, the first few terms give an effective low-energy ($\sqrt{s} \ll m_\rho$) Lagrangian.

Therefore, the starting point is a scaled-up chiral perturbation theory in which the pion is replaced with the Higgs. The lightness of the pion mass is due to its identity being the NGB. The Higgs as the NGB can acquire mass if the global symmetry G is explicitly broken. This is achieved by converting the derivative to a covariant derivative (i.e., by gauging), because the gauge interaction induces the quadratic (i.e., mass term) as well as the quartic field configuration radiatively [6]. If the whole or part of the gauged symmetry contains the EW symmetry, one can fulfill the aim to reproduce the SM as a low-energy effective theory.

The challenge, then, is to fulfill the requirement that the new strong force should only appear beyond the energy scale $\sim O(10)$ TeV while keeping the Higgs mass light. Let us rephrase the problem in more technical terms. Major corrections to the Higgs mass diverge quadratically, as was shown in Eqs. (1.39). The correction has the form

$$m_h^2 = \delta m_h^2 \sim \frac{\alpha_t}{4\pi}\Lambda^2 \tag{1.75}$$

A 125-GeV mass would imply $\Lambda \sim 1$ TeV. To solve the little hierarchy problem, we need $\Lambda \gtrsim O(10)$ TeV while keeping m_h around ~ 100 GeV. How can we achieve this? Suppose that we can arrange the prefactor in front of Λ^2 to be not $(\alpha_t/4\pi)$ but $(\alpha_t/4\pi)^2$; that is, if the leading cutoff sensitivity appears not at one-loop but at two-loop order, then the Higgs mass would be

$$m_h^2 \sim \left(\frac{\alpha_t}{4\pi}\right)^2 \Lambda^2 \tag{1.76}$$

and we may obtain $\Lambda \sim 10$ TeV keeping $m_h \sim 100$ GeV. One should remember that this is not a real solution. The problem is just postponed temporarily. It will reappear as we extend the energy beyond the 10-TeV scale. For the moment, however, we will be content with the temporary solution. The real solution, commonly referred to as the *UV completion*, remains to be solved. The essence of the "little Higgs" is to achieve this extra prefactor ($\alpha_t/4\pi$), that is, to eliminate one-loop correction [61].

A solution to this problem has been proposed [62] by incorporating an enlarged symmetry and embedding two parallel global symmetry breaking in such a way as to compensate the two competing corrections with each other so that only a logarithmic divergence appears. This is referred to as *collective symmetry breaking*. To reproduce the SM Higgs, one has to start with a larger group G. The collective symmetry breaking assumes that G breaks down to a subgroup H which contains $SU(2) \times U(1)$ in the SM. The Higgs appears as an NGB of the broken symmetry. But under normal circumstances, the gauge interaction will induce quadratically divergent contributions to the Higgs mass as in the SM. To avoid this, one assumes that G contains a subgroup consisting of two copies of $SU(2) \times U(1)$: $G \supset H_1 \times H_2 = [SU(2)_1 \times U(1)_1] \times [SU(2)_2 \times U(1)_2]$. The trick is to arrange this in such a way that each H_i commutes with a different subgroup $Y \supset [SU(2) \times U(1)]_{\text{SM}}$. When X, a subgroup of G ($G \supset X \supset Y$), is gauged, the Higgs mass is still protected by the global symmetry of H_1 and H_2.

The group structure of the little Higgs is illustrated in Figure 1.13.

Many models have been proposed. As the group structure of realistic models is rather complicated, we discuss a toy model [28] to understand the mechanism of the collective symmetry breaking, which is the essence of the little Higgs model.

Collective Symmetry Breakdown For simplicity, we omit the $U(1)$ part of the SM and start with a global group $G = SU(3)_L \times SU(3)_R$ breaking to $H = SU(2)_L \times SU(2)_R$. We consider two sets of scalar fields in the fundamental representation that transform independently according to

$$\Sigma_L \to \Sigma'_L = e^{-i\alpha_L}\Sigma_L, \qquad \Sigma_R \to \Sigma'_R = e^{-i\alpha_R}\Sigma_R \tag{1.77}$$

The primary reason that we need two sets of scalar fields is to compensate the two competing corrections with each other. Another reason is that, by gauging a subgroup of the symmetry, one set of the NGB fields is eaten by the gauge particles and disappear. An example can be seen in Eq. (1.73). There, if the broken

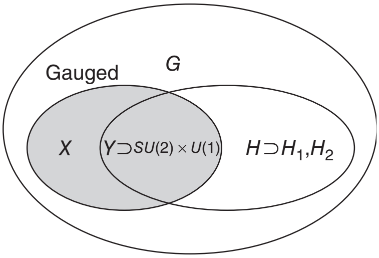


Figure 1.13 Group structure of the little Higgs model. A global group G spontaneously breaks down to H at a scale $f = \Lambda/4\pi$. The origin of the symmetry breaking is irrelevant below the scale Λ. H must contain $SU(2) \times U(1)$ as a subgroup so that, when a part of G, labeled X, is weakly gauged, the unbroken electroweak group $Y = SU(2) \times U(1)$ comes out [61].

symmetry is the gauged $SU(2)$, the NGBs would be absorbed by the gauge bosons of $SU(2)$ and totally disappear. By preparing two sets, at least half of them survive to become the Higgs we want. For compensation, the two scalar fields need to communicate with each other. Therefore, the symmetry to be gauged has to include both contributions equally. It can be achieved by reformulating the two symmetry groups into diagonal groups.

$$\begin{aligned} &SU(3)_L \times SU(3)_R \\ &\quad \to SU(3)_V \times SU(3)_A = SU(3)_{L+R} \times SU(3)_{L-R} \end{aligned} \tag{1.78}$$

The two fields (Σ_L and Σ_R) will transform as **3** under $SU(3)_V$ and as **3** and $\overline{\mathbf{3}}$ under $SU(3)_A$. By gauging $SU(3)_V$, both of them couple to the same gauge fields and thus are able to communicate with each other.

The twofold $SU(3)$ with $8 + 8 = 16$ generators end up with $3 + 3$ unbroken generators corresponding to the two $SU(2)$ groups. This means that $(8 - 3) \times 2 = 10$ generators are broken, thereby yielding 10 massless NGBs. After gauging $SU(3)_V$, five of the NGBs are eaten up, giving mass to five gauge bosons but the other NGBs remain massless at this stage. The onset of the gauge interaction does not break the symmetry at the tree level, but breaks it when higher order terms, that is, radiative corrections, are taken into account. We will come back to this subject shortly.

Denoting the VEV of the symmetry breaking as f, the low-energy (i.e., $E \ll f$) dynamics can be described by two sets of scalars denoted as Σ_1 and Σ_2:

$$\begin{aligned} \Sigma_1(x) &= e^{i\theta_1 f} \begin{bmatrix} 0 \\ 0 \\ f + \rho_1(x) \end{bmatrix} = e^{i\theta_E/f} e^{i\Phi(x)/f} \begin{bmatrix} 0 \\ 0 \\ f + \rho_1(x) \end{bmatrix} \\ \Sigma_2(x) &= e^{i\theta_2 f} \begin{bmatrix} 0 \\ 0 \\ f + \rho_2(x) \end{bmatrix} = e^{i\theta_E/f} e^{-i\Phi(x)/f} \begin{bmatrix} 0 \\ 0 \\ f + \rho_2(x) \end{bmatrix} \end{aligned} \tag{1.79}$$

$$\begin{aligned} \Phi(x) &= \sum_{a=1}^{5} h_a(x) t_{3+a} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & h^+ \\ 0 & 0 & h^0 \\ h^- & h^{0*} & 0 \end{bmatrix} + \frac{\eta}{2\sqrt{3}} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \\ h^\pm &= \frac{h_1 \mp i h_2}{\sqrt{2}}, \quad h^0 = \frac{h_3 - i h_4}{\sqrt{2}}, \quad \eta = h_5 \end{aligned} \tag{1.80}$$

$\rho_{1,2}$ are real scalar fields that have condensed to acquire the VEV f. They have heavy mass $\sim f$ and are integrated out in the effective field theory.[15] The above choice of VEV leaves the $SU(2)_L \times SU(2)_R$ part of the symmetry unbroken. The phase fields θ_E are eaten up by the gauge fields, or are gauged away in mathematical terms. That is, by suitable gauge transformation they are removed from the NGB sector to become third components of the gauge bosons. $t_a = \lambda_a/2$, $a = 4 - 8$ are $SU(3)$ broken generators and λ_a are the Gell–Mann matrices. One sees that the complex scalar $H = \begin{pmatrix} h^+ \\ h^0 \end{pmatrix}$ forms an $SU(2)$ doublet and the real η a singlet. That the field

---

[15] This is a fancy expression to say that at low energies contributions of $\rho_{1,2}$ are negligible.

Φ satisfies the $SU(2)$ symmetry can be seen as follows: Denoting U_2 as $SU(3)$ transformation matrix which conserves its $SU(2)$ part

$$U_2 = \begin{bmatrix} \hat{U}_2 & 0 \\ 0 & 1 \end{bmatrix} \tag{1.81}$$

it is easy to show that Φ obeys the usual $SU(2)$ transformation law $\Phi' = \hat{U}_2 \Phi \hat{U}_2^\dagger$.

$$\begin{aligned} \Sigma \to \Sigma' &= U_2 \Sigma = U_2 e^{i\theta_E/f} e^{i\Phi/f} U_2^\dagger U_2 \begin{bmatrix} 0 \\ 0 \\ f+\rho_1 \end{bmatrix} = e^{i\theta'_E/f} e^{i\Phi'/f} \begin{bmatrix} 0 \\ 0 \\ f+\rho_1 \end{bmatrix} \\ \Phi' &= U_2 \Phi U_2^\dagger = U_2 \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & h^+ \\ 0 & 0 & h^0 \\ h^- & h^{0*} & 0 \end{bmatrix} U_2^\dagger + U_2 \frac{\eta}{2\sqrt{3}} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix} U_2^\dagger \\ &= \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & \hat{U}_2 H \\ H^\dagger \hat{U}_2^\dagger & 0 \end{bmatrix} + \frac{\eta}{2\sqrt{3}} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \end{aligned} \tag{1.82}$$

The symmetry breaking has happened at the scale f, which is higher than the EW symmetry breaking scale, that is, $f \gg v_{\text{EW}} = 246$ GeV. $H(x)$ will become the Higgs field of the SM, but at this stage they are simply massless NGBs.

The effective Lagrangian is expressed in terms of field derivatives because of the shift symmetry, as shown in Eq. (1.73) and (1.74). Note that the shift symmetry not only prohibits the mass term but also forbids the gauge couplings as well as the Yukawa couplings. The identity $\Sigma^\dagger \Sigma = f^2$ constrains the number of independent operators that can be written at each order in the derivative expansion of the Lagrangian. The leading term contains only one term, $\mathscr{L}_2 = \partial_\mu \Sigma^\dagger \partial^\mu \Sigma$, and contains no mass term. By gauging the $SU(3)_V$ part, the dominant Lagrangian is converted to

$$\begin{aligned} \mathscr{L} &= (D_\mu \Sigma_1)^\dagger (D^\mu \Sigma_1) + (D_\mu \Sigma_2)^\dagger (D^\mu \Sigma_2) \\ D_\mu &= \partial_\mu + igA_\mu^a t_a \, (a = 1 \sim 8) \\ \text{where} \quad \Sigma_1 &= e^{i\Phi/f} \begin{bmatrix} 0 \\ 0 \\ f \end{bmatrix}, \qquad \Sigma_2 = e^{-i\Phi/f} \begin{bmatrix} 0 \\ 0 \\ f \end{bmatrix} \end{aligned} \tag{1.83}$$

The Lagrangian Eq. (1.83) describes an effective theory valid only at low energies and is unrenormalizable. The cutoff energy Λ and the symmetry breaking VEV f is related by $\Lambda \simeq 4\pi f$. A simple way to see this is to look at the analogous situation in the SM where the radiative correction of the Higgs loop to the mass is given by $\Delta m_h^2 \simeq \lambda^2 (\Lambda/4\pi)^2$ [see Eq. (1.39c)]. A requirement that it should not exceed the tree level Higgs mass $m_h^2 = 2\lambda v^2$ gives $\Lambda \simeq 4\pi v$.

The Gauge Interaction The gauge interaction induces radiative corrections to the mass and connects the two fields by diagrams described in Figure 1.14. Out of eight gauge bosons, three (the $SU(2)$ part) remain massless, which we call W_L, and

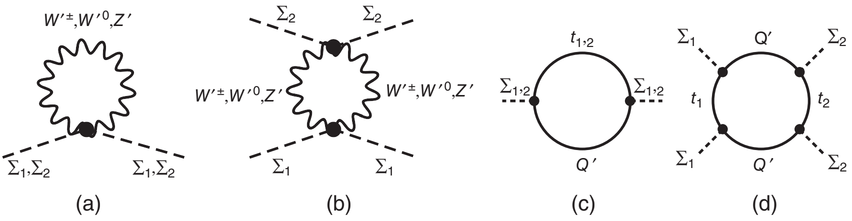


Figure 1.14 Radiative corrections to the Higgs potential. Gauge corrections (a) yield quadratic cutoff dependence which does not contribute to the Higgs potential because W, Z and W', Z' contributions cancel each other, (b) yields log-divergent contribution to the Higgs mass. The Yukawa corrections (c) yields quadratic cutoff dependence which does not contribute to the Higgs potential because t and T contributions cancel each other, (d) yields the log-divergent contribution. $Q' = (t_L, b_L, T_L)$, $t_{1,2} = (T_R \pm it_R)/\sqrt{2}$.

will become the SM gauge bosons later. The remaining five will acquire masses of order f and constitute a $SU(2)$ complex doublet (W'^+, W'^0) and a singlet Z' (to be mixed with yet to enter $U(1)$ group).

After cutting off the loop integral at Λ, Figure 1.14 produces a Higgs potential in the Lagrangian

$$-\frac{g^2}{16\pi^2}\Lambda^2(\Sigma_1^\dagger\Sigma_1 + \Sigma_2^\dagger\Sigma_2) \tag{1.84}$$

As $\Sigma_i^\dagger\Sigma_i = f^2 =$ constant, it does not produce anything. The shift symmetry of Φ is at work here. From the symmetry point of view, the field H contained in Φ is rotated away because of the $SU(2)$ invariant form of $\Sigma^\dagger\Sigma$ which guarantees vanishing mass of the NGBs. For the mass generation of Φ, the Lagrangian has to have some symmetry-breaking term. Let us look at Figure 1.14b. This produces a term

$$-\frac{g^4}{16\pi^2}\ln\left(\frac{\Lambda^2}{f^2}\right)\left|\Sigma_1^\dagger\Sigma_2\right|^2 \tag{1.85}$$

The quadratic divergence terms are canceled between the massless and the massive gauge bosons. So if $\Sigma_1^\dagger\Sigma_2$ constrains the quadratic $H^\dagger H$ terms, we have fulfilled what we had aimed. Expanding Σ_1 and Σ_2 in terms of Φ, we have

$$\begin{aligned}
\Sigma_1 &= e^{i\Phi/f}\begin{bmatrix}0\\0\\f\end{bmatrix} = \left(1 + i\frac{\Phi}{f} - \frac{\Phi^2}{2f^2} + \cdots\right)\begin{bmatrix}0\\0\\f\end{bmatrix} = \begin{pmatrix} ih^+/\sqrt{2}\\ ih^0/\sqrt{2}\\ f\left(1 - \frac{H^\dagger H}{4f^2}\right)\end{pmatrix} + \cdots\\
\Sigma_2 &= \begin{pmatrix} -ih^+/\sqrt{2}\\ -ih^0/\sqrt{2}\\ f\left(1 - \frac{H^\dagger H}{4f^2}\right)\end{pmatrix} + \cdots
\end{aligned} \tag{1.86}$$

where the ellipses $\cdots$ contain higher order terms as well as η-dependent terms. Then

$$\Sigma_1^\dagger \Sigma_2 = -\frac{1}{2}(H^\dagger H) + f^2\left(1 - \frac{H^\dagger H}{4f^2}\right)^2 + \cdots = f^2 - (H^\dagger H) + \frac{(H^\dagger H)^2}{16f^2} + \cdots$$

$$\therefore \quad \left|\Sigma_1^\dagger \Sigma_2\right|^2 = -2f^2(H^\dagger H) + \frac{9}{8}(H^\dagger H)^2 + \cdots \tag{1.87}$$

Thus we have produced a bilinear term as well as a quartic term of the scalar potential that is necessary to reproduce the SM. Notice that the sign of the bilinear term is negative and that of the quartic term is positive as is required for the symmetry breaking and the stability of the potential.[16] After the symmetry breaking, the Higgs mass is given by

$$m_h^2 \simeq \frac{g^4}{16\pi^2} f^2 \ln\left(\frac{\Lambda^2}{f^2}\right) \sim O\left[\left(\frac{f}{4\pi}\right)^2\right] \tag{1.88}$$

where the last equality follows from $\Lambda \simeq 4\pi f$, $g^4 \ln(4\pi)^2 \simeq 1$. As $m_h \sim O(100)$ GeV, we have $f \simeq 4\pi m_h \sim 1$ TeV, $\Lambda \simeq 4\pi f \sim 10$ TeV. Thus the scale Λ of new physics has been pushed off to the safety zone as required by the EW precision data keeping the Higgs mass light.

Yukawa Interaction So far, we have discussed only the gauge field correction to the mass. Now we want to investigate the contribution of the Yukawa, that is, the top quark, interaction. As the original symmetry is $SU(3)$, the fermion in the fundamental representation contains a new fermion field, which we denote as T. Let us consider a left-handed $SU(3)$ triplet $Q'^T_L = (t_L, b_L, T_L)$ and singlets t_R, b_R, T_R. The top quark Yukawa interaction in the SM gives a dangerously large quadratically divergent contribution to the Higgs mass term as we saw in Eq. (1.39a)]. To achieve the same collective symmetry breakdown as the gauge interaction, we expect that contributions of the top quark and the new quark T will compensate each other. Contributions of the other quarks are negligible. Their mass is much smaller.

---

16) Both coefficients of the quadratic and quartic terms are suppressed simultaneously in this simple model. However, a realistic model should have suppression on the bilinear term but not on the quartic term. If both terms are suppressed, it is not possible to simultaneously obtain the correct W boson and phenomenologically acceptable Higgs mass. In the more realistic model, this is achieved by enlarging G to $[SU(4)]^4$, which breaks to $H = [SU(3)]^4$, but this is a complication we will not enter into.

When the symmetry breaks down to $SU(2)$ by the scalar VEVs, the part $Q_L^T = (t_L, b_L)$ inside Q'_L transforms as a doublet under the $SU(2)$. With the following $SU(3)$-invariant Yukawa interaction Lagrangian

$$\begin{aligned} \mathscr{L}_{\text{Yukawa}} &= \frac{h_t}{\sqrt{2}}\left[t_1^c\,\Sigma_1^\dagger Q_L' + t_2^c\,\Sigma_2^\dagger Q_L'\right] \\ h_t &= h_t^{(1)} = h_t^{(2)}, \qquad t_{1,2} = \frac{1}{\sqrt{2}}(T_R \pm i t_R) \end{aligned} \tag{1.89}$$

where $h_t^{(i)}$s are the top Yukawa coupling constants, one can show [61] that the diagrams shown in Figure 1.14c, d exactly yield contributions as given in Eq. (1.84) and Eq. (1.85) in which g is replaced with h_t. Hence, Figure 1.14c gives a quadratically divergent integral for each t and T but cancel each other. Figure 1.14d gives only a logarithmically divergent integral. Thus the collective symmetry breakdown cancellation mechanism is also at work for the Yukawa interaction.

To construct a viable model, we have to extend the symmetry to include $U(1)$ and formulate the whole group structure in such a way as to satisfy phenomenological constraints such as the absence of FCNC. Several models have been proposed.

The most popular model is the littlest Higgs model [63], which is also the most economical in group structure. Its choice of the group is $G = SU(5)$, which breaks to $H = SO(5)$. The subgroup of $SU(5)$ that is gauged is $[SU(2) \times U(1)]_1 \times [SU(2) \times U(1)]_2$, which breaks to $SU(2)_D \times U(1)_Y$. Out of the 14 (=24-10) NGBs generated as a result of $G \to H$ breakdown, 4 are absorbed by the massive $A_H, Z_H, W_H^\pm$ corresponding to the broken $SU(2) \times U(1)$ generators. The other 10 scalars arrange themselves to form a complex $SU(2)$ doublet H with the right quantum number of the SM Higgs plus a complex $SU(2)$ triplet.

Others include the simplest model [64], the minimal moose [65], etc. For more details, one may refer to reviews [28, 58].

After the formulation of the models, construction of a mathematically consistent and closed framework (i.e., UV completion) remains to be solved.

Finally, notice that in the supersymmetric model, cancellation of the divergence was done between particles with different spins. In the little Higgs model, it is done between particles with the same spin.

Experimental signals for the little Higgs are the existence of heavy $W'^\pm$, Z' and the new fermion T. It is a common feature of many beyond-the-SM models and will be discussed in Section 3.4.2.

1.7
Production and Detection of Higgs

Although discovery of the Higgs was at the core of the past effort in going beyond the SM, one should not forget that the most important aim is to understand the Higgs mechanism and not just the discovery of the Higgs particle *per se*. In elucidating the Higgs mechanism, we should consider strategies to cover a broader range of dynamic reactions that are related to the Higgs associated interactions. The Higgs was discovered at LHC. But it is expected that the detailed dynamical structure could only be clarified by an electron collider, that is, the International Linear Collider (ILC).17) Therefore it is important to understand the methodology of detecting the Higgs at the e^-e^+ collider as well as at the hadron collider. We should be aware of the advantage and disadvantage of the hadron collider compared to the electron collider. In view of this, we first review past LEP experiments for the Higgs search before discussing the detection method at the hadron collider.

1.7.1
Higgsstrahlung $e^-e^+ \to hZ$

At LEP, the following two production mechanisms were effective (see Figure 1.15)

$$\bullet\ \text{At}\sqrt{s} = m_Z\ :\ e^-e^+ \to Z \to hZ^* \to hl\bar{l} \qquad (Z^*\text{is virtual}). \qquad (1.90)$$

$$\bullet\ \text{At}\sqrt{s} > m_Z\ :\ e^-e^+ \to Z^* \to hZ \qquad (\text{Higgsstrahlung}). \qquad (1.91)$$

As $m_h > m_Z$, the process (1.90) is no longer relevant for the future consideration. Therefore, we only consider the Higgsstrahlung. At LEPII ($\sqrt{s} = 200$ GeV), the method to use the Higgsstrahlung [Figure 1.15a] was viable. The selection of events was made by looking at $Z \to l\bar{l},\ q\bar{q}$, requiring the invariant mass of the lepton or quark pair to coincide with m_Z. Identification of $Z \to \nu\bar{\nu},\ h \to q\bar{q}$ was also possible by requiring "1 jet + missing energy." This is possible because the total energy of the Zh system is known and the Higgs mass can be reconstructed from the observables. The Higgs was not discovered at LEP, and the obtained upper limit was 114 GeV at 95% CL.

1.7.2
W Boson Fusion

When the Higgs mass goes beyond 100 GeV, the vector boson fusion process ($e^-e^+ \to e^-e^+V^*V^* \to e^-e^+h,\ V = W, Z$) becomes the dominant process in the e^-e^+ collider (see Figure 1.15b, c). Although the gluon fusion is the dominant process at LHC as far as Higgs production is concerned, the W boson fusion will again take over for sufficiently large s. The WW reaction has a unique feature that the gluon fusion does not have. Longitudinal components of W are the would-be-Goldstone bosons of the Higgs, that is, it is a direct result of the spontaneous symmetry breaking. In fact, there exists an equivalence theorem that, at sufficiently high

---

17) The ILC is considered as the next major project in particle physics. It is an electron–positron collider using straight linacs stretching up to 20 km in length. The energy has not been decided yet but is envisaged to start from ~500 GeV eventually going over 1 TeV. It could be set lower to optimize for the Higgs study.

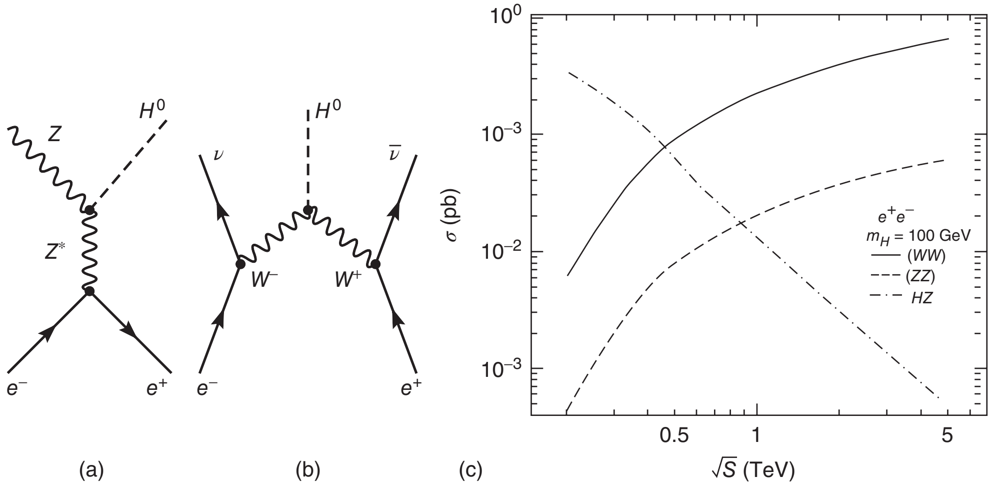


Figure 1.15 Higgs production mechanisms and their total cross section. (a) Higgsstrahlung: $e^-e^+ \to Z^* \to Zh$. Valid for $m_Z < \sqrt{s} \lesssim 300$ GeV and 60 GeV$< m_h < 100$ GeV. (b) W boson fusion: $e^-e^+ \to W^-W^+ \to h$. Valid for $\sqrt{s} > 300$ GeV and $m_h > 100$ GeV. (c) Total cross section for the Higgs production via vector boson fusion in e^-e^+ annihilation as a function of $\sqrt{s}$ when $m_h = 100$ GeV. The total cross section for the process $e^-e^+ \to hZ$ is also shown for comparison. (Reproduced with permission of [66].)

energy ($s \gg m_W^2$), W_LW_L, W_LZ_L scattering are equivalent to the scatterings between the would-be-Goldstone bosons (h^0, $h^\pm$) [67]. In this regard, the WW reactions are inherently suited to elucidate the dynamics of the Higgs mechanism.

In order to understand the WW reaction mechanism, including the Higgs production at the hadron collider, we first study the idea of boson fusion and the equivalent W approximation in the electron collider. In the hadron collider, the main QCD processes are not the quark–quark interactions but the gluon–gluon reaction. It is related partly to the large coupling of the gluon but also has its origin in the vanishing mass of the gluon. In the EW production of the Higgs at the hadron collider, the total energy is large enough so that $\hat{s} \gg m_W^2$, where $\hat{s}$ is the total energy of the elementary process of interest (i.e., partons). At such high energies, even the gauge bosons can be regarded as nearly massless and we expect the boson–boson interactions to play an important role just like the gluon–gluon interactions in QCD. In this respect, we can make use of the tools we acquired in quantum electrodynamics (QED). The following discussion is provided to prepare the WW scattering as one of main tools for probing the dynamics of the Higgs sector after the discovery of the Higgs particle.

Equivalent *W* Approximation As for the boson fusion cross section, an exact formula in the tree approximation exists [66, 68], but the effective W approximation, which is easier to understand intuitively, gives a good approximation (error < 20%) and we will use it hereafter. This is an application of the Weiszäker–William approximation (see Section 17.6.3 of [1]). It replaces an electromagnetic e–A (A represents an atom) process by the equivalent γ^*–A process where the almost-real photon flux is provided by the electron. Replacing the photon with W and the electromagnetic coupling by the EW coupling, one gets the equivalent W approximation. In other words, the equivalent W approximation is valid at the high energy where the relevant total energy $\hat{s}$ is sufficiently large compared to the W mass. Then the Bremsstrahlung formula in QED can be used, which almost restricts the W emission in the forward region ($\theta \approx 0$). One may recall that a similar consideration was adopted in formulating the DGLAP (Dokshitzer–Gribov–Lipatov–Altarelli–Parisi) evolution equation in QCD to compute the parton flux. Here, the electron is replaced by the quark, and the virtual photon by the gluon.

In phenomenological expressions, the W mass is retained in the propagator, which provides a natural cutoff below $p_T < m_W$. In this approximation, we consider the incoming e^-/e^+ beam as a supplier of the gauge boson flux. Once we get an equivalent flux function, it can be treated like the parton distribution functions in QCD, and the WW, ZZ, ZW scattering formula can be constructed using factorization formulas as used in hadron–hadron collisions.

Let us start from the W bremsstrahlung by an electron. Extension to that by a quark is straightforward. Assume that the electron with energy-momentum p_1 and the positron with p_2 are the suppliers of the colliding Ws with momenta $x_1 p_1$ and $x_2 p_2$ of the parents e^-e^+ and that its flux is given by $F(x_1)$ and $F(x_2)$. The cross section for the Higgs production $\sigma(e^-e^+ \to e^-e^+h; s)$ can be expressed in terms of the vector boson fusion process cross section $\sigma(VV \to h; \hat{s})$:

$$\begin{aligned}\sigma(e^-e^+ \to e^-e^+h; s) &= \int dx_1 dx_2\, F(x_1)F(x_2)\hat{\sigma}(VV \to h; \hat{s}) \\ \hat{s} &= x_1 x_2 s\end{aligned} \tag{1.92}$$

For $m_h^2 \gg m_W^2$ and $\tau = m_h^2/s$

$$\begin{aligned}\hat{\sigma}(VV \to h) &= \frac{16\pi^2 m_h}{\hat{s}}\Gamma(h \to VV)\delta(\hat{s} - m_h^2) \\ &= \frac{16\pi^2 m_h}{\hat{s}^2}\Gamma(h \to VV)\tau\delta(x_1 x_2 - \tau)\end{aligned} \tag{1.93}$$

Substituting Eq. (1.93) into Eq. (1.92), we obtain

$$\sigma(e^-e^+ \to e^-e^+h) = \frac{16\pi^2 m_h}{\hat{s}^2}\Gamma(h \to VV)\tau\frac{dL}{d\tau} \tag{1.94}$$

where

$$\frac{dL}{d\tau} = \int_\tau^1 \frac{dx}{x}F(x)F(\frac{\tau}{x}) \tag{1.95}$$

$dL/d\tau$ is the luminosity of the WW flux that the e^-e^+ beam provides. Substituting the expression for the decay rate [Eqs. (1.19)] into the above expression, we obtain

$$\sigma(e^-e^+ \to e^-e^+h) = \frac{\pi^2\alpha}{\sin^2\theta_W m_W^2}\tau\frac{dL}{d\tau} \tag{1.96}$$

Kinematics of the W flux in $e \to e + W$ prepared by its parent particle is the same as that of the gluon flux in QCD with replacements of the coupling constant and symmetry factors. Thus the flux of W prepared by the electron can be obtained using the same formula as used to derive the splitting functions in QCD.

$$\begin{aligned} F(x) &= \frac{g^2}{8\pi^2}P_{\mathrm{BC}\leftarrow A}(x) \\ P_{\mathrm{BC}\leftarrow A}(x) &= \frac{x(1-x)}{2}\frac{\overline{\sum}|V(A\to \mathrm{BC})|^2}{p_T^2+(1-x)m_W^2} \quad ^{18)} \end{aligned} \tag{1.97}$$

where g is the coupling constant of W/Z with the electron. Despite the inherent zero-mass approximation for the W bremsstrahlung, the mass term in the denominator was retained to provide a natural cutoff at small p_T. Substituting actual expressions of the $(e \to e + W)$ Lagrangian, one gets

$$g^2\overline{\sum}|V(A\to BC)|^2 = \overline{\sum}|\overline{u}(p_C)\not{\epsilon}(g_V - g_A\gamma^5)u(p_A)|^2 \tag{1.99}$$

Here, g_V and g_A are the vector and the axial-vector coupling strength and are different for $V = W$ or $V = Z$. Using

$$\begin{aligned} W: &\quad g_V = g_A = \frac{g_W}{2\sqrt{2}} \\ Z: &\quad g_V = \frac{g_Z}{2}\left(I_3 - 2Q\sin^2\theta_W\right), \quad g_A = \frac{g_Z}{2}I_3 \\ &\quad g_W = \frac{e}{\sin\theta_W}, \qquad g_Z = \frac{e}{\sin\theta_W\cos\theta_W} \end{aligned} \tag{1.100}$$

---

18) This is a general formula [69] to calculate the splitting function in QCD when a parton "a" splits into partons "b" and "c," the former having fractional momentum $p_b = xp_a$ and the latter $p_c = (1-x)p_a$, except that in QCD the mass is set to zero.

$$\begin{aligned} &d\sigma(a + d \to c + X) \\ &\simeq F_{bc\leftarrow a}(x)dx\frac{d\phi}{2\pi}d\ln p_T^2 d\sigma(b + d \to X) \end{aligned}$$

$$F_{bc\leftarrow a} = \frac{\alpha_s}{2\pi}\left[\frac{x(1-x)}{2}\frac{|\overline{\sum}|V(a\to bc)|^2}{p_T^2}\right] \tag{1.98}$$

See Appendix K of [2].

one calculates the flux for $\hat{s} \gg m_W^2$ to obtain [70, 71]

$$F_T(x) = \frac{(g_V^2 + g_A^2)}{8\pi^2} \frac{1 + (1-x)^2}{x} \ln\left(\frac{\hat{s}}{m_W^2}\right) \tag{1.101a}$$

$$F_L(x) = \frac{(g_V^2 + g_A^2)}{4\pi^2} \frac{(1-x)}{x} \tag{1.101b}$$

The luminosity function $F_T(x)$ for the transversely polarized W is identical to the Weiszäcker–Williams formula (see Eq. (17.89) of [1]) for the photon flux given by the electrons if one replaces $(g_V^2 + g_A^2)/(4\pi) \to \alpha$. Substituting Eqs. (1.101) in Eq. (1.95), one gets the VV luminosity for the transversely and longitudinally polarized gauge bosons [72].

$$\left.\frac{dL}{d\tau}\right|_{ee/V_T V_T} = \left[\frac{g_V^2 + g_A^2}{8\pi^2}\right]^2 \left[\ln \frac{\hat{s}}{m_W^2}\right]^2 \frac{1}{\tau}\left[(2+\tau)^2 \ln\frac{1}{\tau} - 2(1-\tau)(3+\tau)\right] \tag{1.102a}$$

$$\left.\frac{dL}{d\tau}\right|_{ee/V_L V_L} = \left[\frac{g_V^2 + g_A^2}{4\pi^2}\right]^2 \frac{1}{\tau}\left[(1+\tau)\ln\frac{1}{\tau} - 2(1-\tau)\right] \tag{1.102b}$$

The Higgs production cross section becomes [66, 68, 72]

$$\begin{aligned}\sigma(e^- e^+ \to e^- e^+ h) = \left(\frac{\alpha}{\sin^2\theta_W}\right)^3 \frac{1}{16 m_h^3}\Bigg[\frac{m_h^3}{m_W^2}&\left\{(1+\tau)\ln\frac{1}{\tau} - 2(1-\tau)\right\} \\ + \frac{m_W^2}{2m_h}\left(\ln\frac{\hat{s}}{m_W^2}\right)^2 &\left\{(2+\tau)^2\ln\frac{1}{\tau} - 2(1-\tau)(3+\tau)\right\}\Bigg]\end{aligned} \tag{1.103}$$

The first line is the contribution of the longitudinal W, and the second line is that of the transverse W's.

1.7.3
Productions at the Hadron Collider

The main mechanism of the Higgs production at the LHC is gluon fusion, and the submechanism is vector boson fusion. Their Feynman diagrams are shown in Figure 1.16a,c. Two other reactions that can be used to identify the Higgs are $t\bar{t}h$-associated production (Figure 1.16d) and Higgsstrahlung by the quarks (Figure 1.16e).

Gluon Fusion The production cross section of the Higgs by gluon fusion is given by [73, 74]

$$\begin{aligned}\sigma(pp \to pph) &= 2 \times \frac{1}{4} \times \frac{1}{64} \times \frac{16\pi^2}{m_h^3}\Gamma(h \to gg)\,\tau\frac{dL}{d\tau} \\ \frac{dL}{d\tau} &= \int dx_1 dx_2\, \delta(x_1 x_2 - \tau) g(x_1) g(x_2)\end{aligned} \tag{1.104}$$

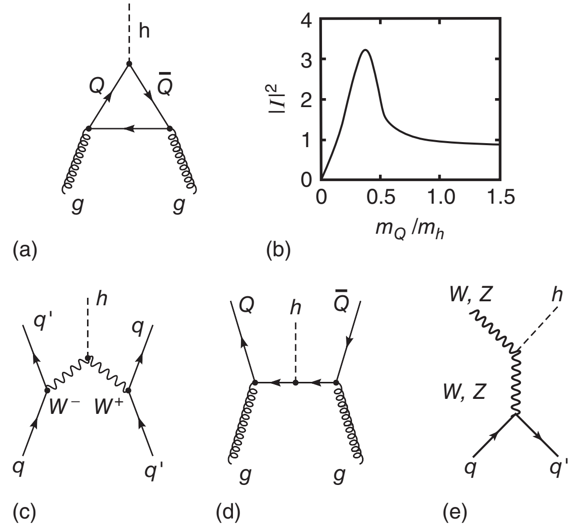


Figure 1.16 Feynman diagrams for the Higgs production. (a) Gluon fusion. (b) Shape of the loop integral of the gluon fusion diagram. (c) Boson fusion. (d) $t\bar{t}h$-associated production. (d) QCD + EW production

where $g(x)$ is the gluon distribution function in the proton. The extra factors in front of $\Gamma(h \to gg)$ relative to Eq. (1.94) (1/4, 1/64) are due to spin and color degrees of freedom of the gluon, and the factor 2 is due to the Bose–Einstein statistics of the two gluons. The decay width of $h \to gg$

$$\Gamma(h \to gg) = \frac{\sqrt{2}G_F[\alpha_s(m_h)]^2}{8\pi^3}\frac{m_h^3}{9}|I|^2 \tag{1.105}$$

can be derived from an effective Lagrangian

$$\mathscr{L}_{hgg} = -\left(\sqrt{2}G_F\right)^{\frac{1}{2}}\frac{\alpha_s(m_h)}{12\pi}IG_{\mu\nu}G^{\mu\nu}h \tag{1.106}$$

where $G_{\mu\nu}$ is the gluon field strength. Here, the quantity I originates from fermion loops and is given by

$$I = \sum_j I_j = N_c \sum_j \int_0^1 dx \int_0^{1-x} dy \frac{1-4xy}{1-xy\left(\frac{m_h^2}{m_j^2}\right)-i\varepsilon} \tag{1.107}$$

The spectral shape of the loop integral for the gluon fusion $|I|^2$ is plotted in Figure 1.16b. Its value is dominated by the heaviest quark, that is, the top. $|I|^2$ takes a maximum value 3.2 for $\lambda_Q = m_Q/m_h \simeq 0.4$ and $|I|^2 \simeq 1$ for $\lambda_Q \gtrsim 1$. For $\lambda_Q \ll 1$, $|I|^2 \sim (\lambda_Q \log \lambda_Q)^4$. That is, $\sigma(gg \to h)$ decreases like $\lambda_Q^4 \sim 1/m_h^4$. The total production cross section of the Higgs in the hadron collider which incorporates NNLO in QCD [75] is given in Figure 1.17. The largest cross section comes from $\sigma(gg \to h)$, which is the top line. The broad bump in the cross section reflects the shape of the loop integral.

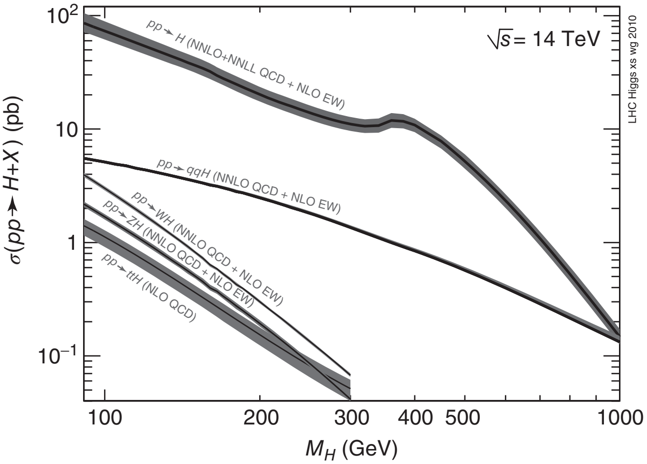


Figure 1.17 SM Higgs boson production cross sections for pp collisions at 14 TeV. The top line is due to the gluon fusion, and the line $pp \to qqH$ denotes the W boson fusion. Other channels are also indicated. (Reproduced with permission of [8, 76, 77].)

Luminosity of *W* Boson Fusion The W flux provided by a quark is identical to that by an electron except for replacement of $e - W$ coupling with $q - W$ coupling. The quark flux, in turn, is provided by the proton in a hadron collider. Therefore, the Higgs production through WW fusion in the hadron collider can be obtained from that of the electron collider by convoluting with parton distribution functions. Or, equivalently, the W luminosity in the pp reaction is given by convoluting that of e^-e^+ collision with the parton distribution functions $f_i(x)$ [70, 71].

$$\left.\frac{dL}{d\tau}\right|_{pp/VV} = \sum_{ij}\int dx_1 dx_2 f_i(x_1) f_j(x_2) \times \left[\int d\xi d\eta\, F(\xi)F(\eta)\delta(x_1\xi x_2\eta - \tau)\right] \tag{1.108}$$

Putting $x_1x_2 = \tau'$, the content of $\left[\cdots\right]$ is expressed as

$$\frac{1}{\tau'}\int \frac{dx}{x}F(x)F\left(\frac{\frac{\tau}{\tau'}}{x}\right) = \frac{1}{\tau'}\left.\frac{dL}{d\xi}\right|_{\xi=\tau/\tau'} \equiv \frac{1}{\tau'}\left.\frac{dL}{d\xi}\right|_{qq/VV} \tag{1.109}$$

The WW luminosity in the pp collision becomes

$$\left.\frac{dL}{d\xi}\right|_{pp/VV} = \sum_{ij}\int_{\tau}^{1}\frac{d\tau'}{\tau'}\int_{\tau'}^{1}\frac{dx}{x}f_i(x)f_j\left(\frac{\tau'}{x}\right)\left.\frac{dL}{d\xi}\right|_{qq/VV} \tag{1.110}$$

Then the cross section is given by

$$\sigma(pp \to VV \to h; s) = \int_{\tau}^{1} d\tau \left.\frac{dL}{d\tau}\right|_{pp/VV}\hat{\sigma}(VV \to h; \tau s) \tag{1.111}$$

where the variables in $\hat{\sigma}(VV \to h)$, $dL/d\xi|_{qq/VV}$ are to be replaced from those of the electron collision to those of the quark collision. The reduction rate of the W boson fusion cross section as a function of m_H is much slower than that of the gluon fusion cross section. So, it will eventually dominate over the gluon fusion, but at the LHC the dominant contribution still comes from the gluon fusion. Detailed calculations at NLO (next-to-leading order) in QCD have been obtained by [78, 79]. Figure 1.17 gives the most recent plot on the cross sections including other channels [8, 77]. The main contributions to the Higgs production cross sections come from gluon or W fusions, but some other modes are also useful for obtaining better S/N (signal-to-noise ratio) by identifying the accompanying particles.

1.7.4
Signals at LHC

We already know that the Higgs was discovered with mass in the neighborhood of 125 GeV. Nevertheless, we discuss possible signals of Higgs production and decays in the various mass regions because the discovery alone is not our final goal. We want to study the dynamics of the Higgs, and it is useful to consider how it would change if the Higgs mass was different. Study of competing processes is also important.

In the e^-e^+ reaction, the competing processes posses physical meanings of their own right. All the open channels have branching ratios of the same order (see Figure 1.18a), and are useful in elucidating one aspect of the physics or another.

Considering channels other than that of specific interest as noises, the S/N in the e^-e^+ reaction is at most of the order of 100. Therefore, if enough number of events are obtained, it is relatively easy to identify the produced Higgs by looking at its decay products such as $h \to jj\,(2\,\text{jets}),\ \bar{l}l\,(l = e, \mu, \tau)$. In the two-body production $e^-e^+ \to Zh$, the invariant mass of the Higgs can be obtained by simply identifying the Z.

But in hadron productions, backgrounds due to QCD processes are large and S/N for the Higgs production is minuscule, as can be seen from Figure 1.18b. To illustrate the difficulty of the event selection in the hadron collider, we show a typical event display at the LHC in Figure 1.19. There are ~25 vertices per beam crossing. The bunch crosses every 25 ns, totaling up to 600 million collisions per second. A signal in such a proliferation of events is literally a needle in a haystack.

One has to choose the decay channels carefully to identify the parent Higgs, depending on its mass.

At LHC, for a luminosity of $L = 10^{34}/\text{cm}^{-2}/\text{s}^{-1}$.[19] one can obtain in a year (assume 1 year = 10^7 s) $\int L = 10^{41}/\text{cm}^{-2} = 100\ \text{fb}^{-1}$. For the process having a total cross section 1 *pb* one can obtain 10^5 Higgs particles. Let us investigate whether this number is sufficient to discover the Higgs. As the decay branching ratio changes as a function of the mass (see Figure 1.2) and backgrounds for each channel are different, an optimum detection method has to be adopted depending on the mass value. As the experimental lower limit had already reached a value $m_h = 114$ GeV before the LHC, we consider cases for a mass larger than this value. The following

---

[19] The improved design luminosity is up to $5 \times 10^{34}/\text{s}^{-1}/\text{cm}^{-2}$. So this is a modest number.

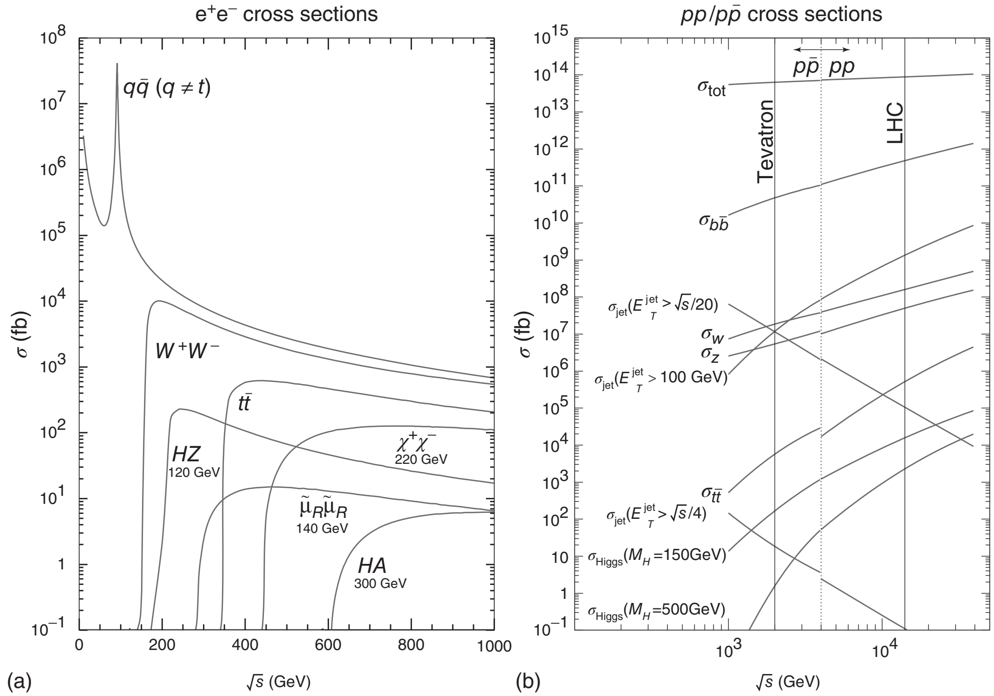


Figure 1.18 Production cross sections for several representative processes at e^+e^- colliders (a) and hadron colliders (b), as a function of the machine center-of-mass energy. (Reproduced with permission of [80].)

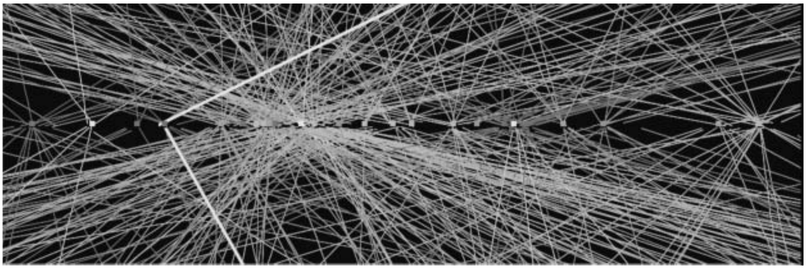


Figure 1.19 Event display for a trigger per beam crossing in the ATLAS detector with 25 reconstructed vertices. This event happened to include a $Z \rightarrow \mu\mu$ event. The display with track p_T threshold of 0.4 GeV and all tracks are required to have at least three pixels and six SCT (SemiConductor Tracker) hits. (Reproduced with permission of [81, 82].)

list is only an example. Although we have listed some useful modes in discovering the Higgs for particular mass range, most of the modes can be used eventually at any mass range as the intensity goes up and backgrounds are better understood. Comparison of observed data with them will provide important information on the properties of the Higgs.

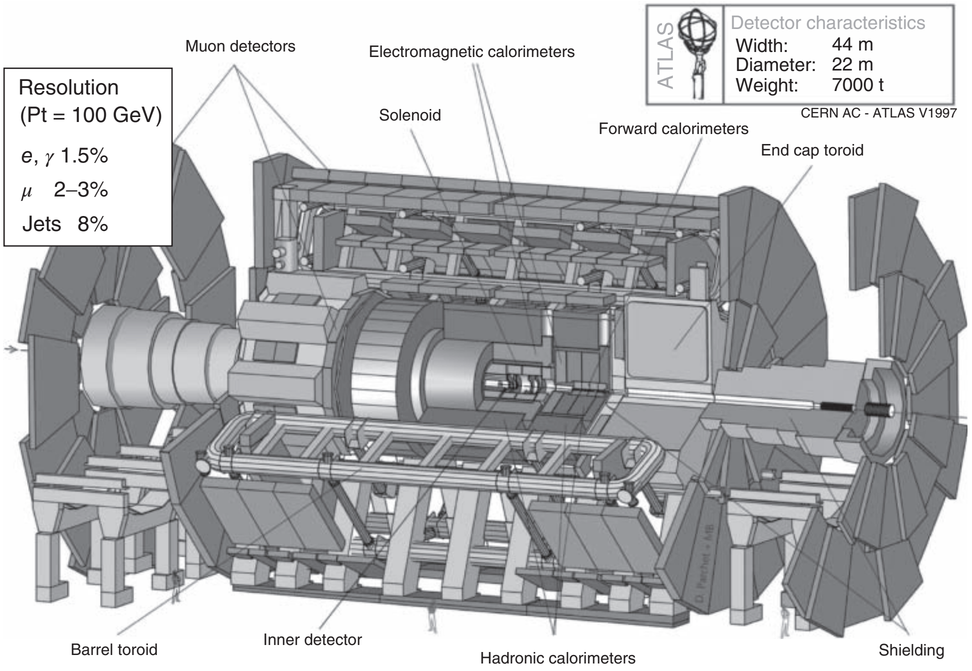

Resolution
(Pt = 100 GeV)
e, γ 1.5%
μ 2–3%
Jets 8%
Muon detectors
Electromagnetic calorimeters
Solenoid
ATLAS
Detector characteristics
Width: 44 m
Diameter: 22 m
Weight: 7000 t
CERN AC - ATLAS V1997
Forward calorimeters
End cap toroid
D. Parchet + MB
Barrel toroid
Inner detector
Hadronic calorimeters
Shielding

Figure 1.20 ATLAS (A Toroidal LHC ApparatuS) is a general-purpose detector with balanced performance including hadronic events. Its magnet is an air-cored one with inner (outer) radius 5 (10)m and length 26 m, and produces a field strength of 0.8 T (2 T at the center). The electromagnetic calorimeter is a liquid argon detector of accordion shape [see Figure 12.35 of [1]]. The overall size is $20\phi \times 44$ m, and weighs 6000 tons. Resolutions are good but not particularly good for all measurements. (Reproduced with permission of [83].) (Please find a color version of this figure on the color plates.)

Two large general-purpose detectors, ATLAS and CMS were constructed and are in operation (see Figure 1.20, 1.21). The CMS has a better resolution for gamma/lepton signals but is less versatile for hadron detection. Components of the general purpose collider detectors and their functions were explained in Section 12.6 of [1]. Two special purpose detectors, LHCb for the B-physics and Alice for the heavy ion collisions were also constructed.

1.7.5
Higgs Detection Methods

(1) 110 GeV< m_h <130 GeV $h \rightarrow \gamma\gamma$: For the mass $m_h < 2m_W$, $h \rightarrow b\bar{b}$ is the main decay mode. However, the QCD background is overwhelming and their separation is difficult. Looking at Figure 1.2, there exists a usable mode $h \rightarrow \gamma\gamma$ with a branching ratio of $\sim 10^{-3}$. That is, in the inclusive decay mode

$$p + p \rightarrow h + X, \qquad h \rightarrow \gamma + \gamma \tag{1.112}$$

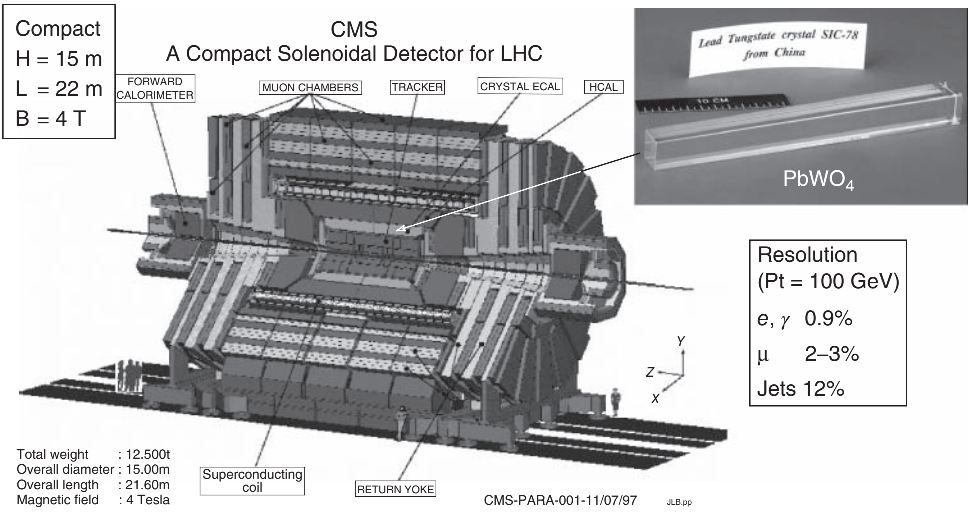

Compact
H = 15 m
L = 22 m
B = 4 T
CMS
A Compact Solenoidal Detector for LHC
FORWARD CALORIMETER
MUON CHAMBERS
TRACKER
CRYSTAL ECAL
HCAL
Lead Tungstate crystal SIC-78 from China
PbWO4
Resolution
(Pt = 100 GeV)
e, γ 0.9%
μ 2–3%
Jets 12%
Y
Z
X
Total weight : 12.500t
Overall diameter : 15.00m
Overall length : 21.60m
Magnetic field : 4 Tesla
Superconducting coil
RETURN YOKE
CMS-PARA-001-11/07/97
JLB.pp

Figure 1.21 CMS (compact muon solenoid) is a detector at LHC. The length is about one-half of that of the ATLAS detector but its weight (= 12.5 kton) is twice that of ATLAS. It uses $PbWO_4$ (lead glass) for the electromagnetic calorimeter and realizes a resolution of 0.9% (1.5%) at $P_T = 100$ GeV. The magnetic field strength is 4 T as compared to 2 T of ATLAS. It excels in $e/\mu/\gamma$ physics and is general purpose, but is less versatile for hadron detection. (Reproduced with permission of [84].) (Please find a color version of this figure on the color plates.)

one catches two γ's and identifies their parent Higgs by reconstructing the 2γ invariant mass. However, catching 2γ alone may suffer from large backgrounds due to 2γ production by QCD and misidentification of QCD jets. One needs a γ detector (electromagnetic calorimeter) with high energy resolution [3,85]. This is a clean channel but needs high luminosity.

$\boldsymbol{t\bar{t}h\,(h \to b\bar{b})}$: Another method is to use associated production accompanied with W or $t\bar{t}$ in $q\bar{q} \to WhX$, $t\bar{t}hX$ and identify isolated leptons from W or $t\bar{t}$. With this method, one can obtain cleaner signals, but the problem is whether one can get enough statistics [86–88]. The experimental groups ATLAS and CMS are capable of discovering the Higgs in this decay mode within the mass range $80 - 130$ GeV [89,90].

(2) 120 GeV< $\mathbf{m_h}$ <180 GeV $\boldsymbol{h \to ZZ^* \to 4l}$: This is a golden mode in the search for the Higgs. Although the mass value is too low to decay to two real Z's, production of four leptons through virtual Z intermediate states is possible. This channel is promising for the range $120 < m_h < 180$ GeV except for the narrow region around 170 GeV (see Figure 1.2).

$\boldsymbol{h \to \tau\bar{\tau}}$ (m_h <140 GeV): One looks for $\mu^{\pm}+$ missing energy + X [91]. The Higgs mass cannot be reconstructed in this channel, but the signal can be observed as an excess of events above backgrounds which come mainly from WWX and $t\bar{t} \to WWb\bar{b} \to l\bar{l}\nu\bar{\nu}$.

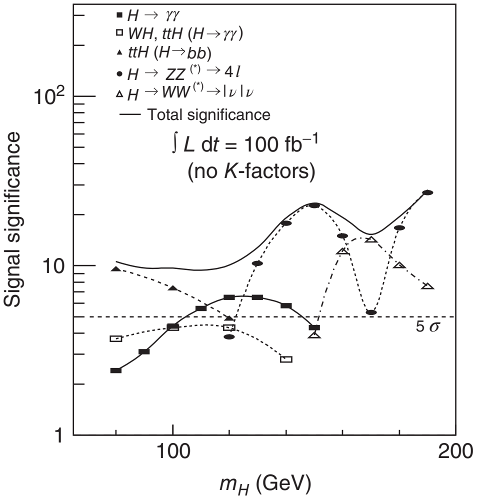


Figure 1.22 ATLAS sensitivity for the discovery of a Standard Model Higgs boson for integrated luminosities of 100 fb^{-1} data in the $m_H < 200$-GeV range. The dotted line is the 5σ significance level. The figure was taken from design studies for the detector construction of the ATLAS. The actual results are much better. (Reproduced with permission of [94, 95].)

$\boldsymbol{h \to W^-W^+ \to l\bar{l}\nu\bar{\nu}}$ ($m_h \sim 170$ GeV) [92, 93]: For the Higgs mass nearing $2m_Z$, the signal significance in the $h \to ZZ^* \to 4l$ channel is reduced because of the suppression of the ZZ^* branching ratio as the WW decay mode opens up. For $m_h = 170\,\text{GeV}$, $h \to WW^* \to l\bar{l}\nu\bar{\nu}$ branching ratio is approximately 100 times larger than that of the $h \to ZZ^* \to 4l$ channel. However, because of the presence of the neutrino, it is not possible to reconstruct the true Higgs mass. Instead, an excess of events may be observed. The transverse mass

$$m_T = \sqrt{2p_T^{ll}E_T^{\text{miss}}(1 - \cos\Delta\phi)} \tag{1.113}$$

is expected to fall in the mass window $m_h - 30\,\text{GeV} < m_T < m_h$. Here, p_T^{ll}, E_T^{miss}, and $\Delta\phi$ are momentum of the lepton pair, the missing transverse energy, and the azimuthal angle in the transverse plane, respectively.

We summarize the Higgs discovery sensitivity of the ATLAS detector below $m_h = 200$ GeV in Figure 1.22 for integrated luminosity of 100 fb^{-1}[20].

(3) 180 GeV< $\boldsymbol{m_h}$ <800 GeV $\boldsymbol{h \to ZZ \to 4l}$: If the Higgs mass is larger than $2m_Z$, the main decay modes are WW and ZZ pairs. Hadronic decay modes are difficult to separate completely from QCD backgrounds. The cleanest signal can be obtained from purely leptonic decay modes

---

[20] Notice that the Higgs was discovered with an integrated luminosity of 5 fb^{-1} at 7 TeV.

$$\begin{aligned} &p + p \to h + X \\ &\qquad h \to Z + Z \to e^- e^+(\mu^- \mu^+) + e^- e^+(\mu^- \mu^+) \end{aligned} \tag{1.114}$$

This decay mode is referred to as the *gold-plated channel* [96–99]. As the branching ratio is 1.1×10^{-3}, the signal size is not large. But by requiring the invariant mass of the two pairs of leptons $m_{\bar{l}l} = m_Z$, one can reduce a large amount of the background. The main components of the background include ZZ continuous spectrum production and $Zb\bar{b}$, $Zt\bar{t}$.

(4) 700 GeV< m_h <1000 GeV $\quad h \to ZZ \to \bar{l}l\nu\bar{\nu}$:

The discovery capability using the $h \to ZZ \to 4l$ decay mode goes down as the mass goes up and hits the limit at around 800 GeV. This is mainly due to the reduction of the counting rate but also because the width becomes large ($\Gamma_h > 250$ GeV). The resonance is no longer a sharp peak but a slow bump and is hard to identify. There is also a theoretical question as to whether such a wide resonance can be considered as a particle.

To extend the mass range, one needs to use other channels that include neutrinos in the final state. For the large-mass Higgs, if the hermeticity of the detector is good (i.e., if a condition that the missing energy vector can be well defined for large rapidity holds), one may be able to reconstruct the missing energy. The Higgs signal in this case appears as the Jacobian peak in the transverse momentum,[21] a fact that is utilized in identifying the W boson at the Tevatron. Because of the facts that one cannot reconstruct the invariant mass and the width becomes large, one may not be able to expect a sharp peak in the event distributions. The signal appears as an excess over the background, and therefore understanding the background is very important for the detection of the Higgs using the missing energy.

In summary, the LHC was designed to discover the SM Higgs having mass $110\,\text{GeV} < m_H < 1\,\text{TeV}$. It covers the entire mass range that is theoretically allowed.

1.7.6
Discovery of Higgs

Both ATLAS and CMS groups announced the discovery of Higgs-like particles at the mass value $m_h \simeq 125$ GeV in the summer of 2012. We present the signals in the decay channels $h \to \gamma\gamma$ in Figure 1.23 and $h \to 4l$ in Figure 1.24.

The Higgs-like boson was produced with approximately the right values for production cross sections as well as for branch decay modes that are expected for the Higgs in the SM.

---

[21] A peak at the edge of a spectrum that appears by a sharp cutoff of kinematic variables (in this case p_T).

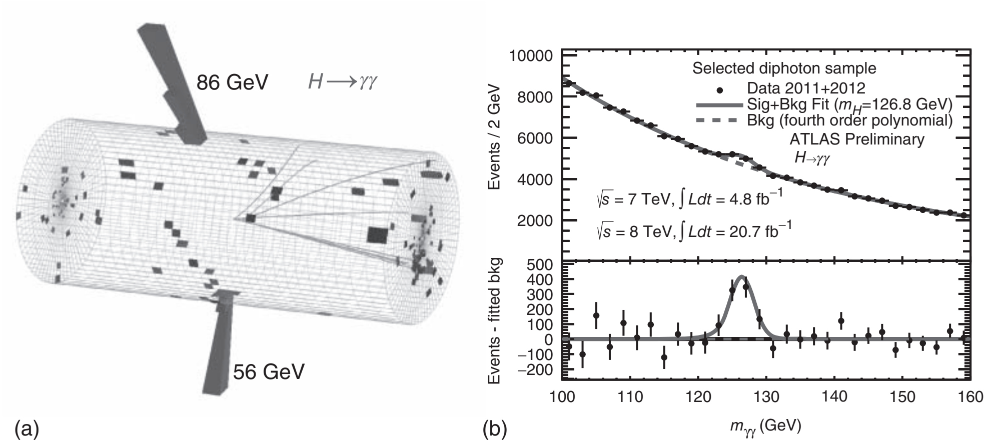


Figure 1.23 (a) CMS event display for the Higgs in the $h \to 2\gamma$ mode. (Reproduced with permission of [100, 101].) (b) Invariant mass distribution of diphoton candidates for the combined $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV data samples obtained by ATLAS. The result of a fit to the data of the sum of a signal component fixed to $m_H = 126.8$ GeV and a background component described by a fourth-order Bernstein polynomial are superimposed. The bottom inset displays the residuals of the data with respect to the fitted background component. (Reproduced with permission of [102].)

1.7.7
SM Higgs?

The question arises whether the discovered Higgs-like particle is really the SM Higgs or something similar to it in other models.

Spin Parity of the Higgs The spin parity of the Higgs-like particle is converging to 0^+, although statistics are poor at this stage (summer, 2013). The CMS group analyzed the spin parity of the Higgs using a kinematic discriminant which includes the description of the interference of identical leptons in the $4e$ and 4μ final states. The discriminant adopts the matrix element likelihood approach (MELA) which uses

$$\mathcal{D}_{J^P} = \frac{\mathcal{P}_{\rm SM}}{\mathcal{P}_{\rm SM} + \mathcal{P}_{J^P}} = \left[1 + \frac{\mathcal{P}_{J^P}(m_{Z_1}, m_{Z_2}, \vec{\Omega}|m_{4l})}{\mathcal{P}_{\rm SM}(m_{Z_1}, m_{Z_2}, \vec{\Omega}|m_{4l})}\right]^{-1} \tag{1.115}$$

as a discriminant between the different spin-parity hypotheses. Here, $\mathcal{P}_{\rm SM}$ is the probability distribution for the SM Higgs boson hypothesis, and $\mathcal{P}_{J^P}$ is the probability for an alternative model. Figure 1.25a shows the distributions of $\mathcal{D}_{J^P}$. Distributions in data (points with error bars) and expectations for background and signal are shown for the alternative hypothesis between 0^+ and 0^-. Figure 1.25b shows a distribution of $q = -2\ln(\mathcal{L}_{J^P}/\mathcal{L}_{\rm SM})$ for the 0^+ and 0^- hypothesis. The arrow indicates the observed value.

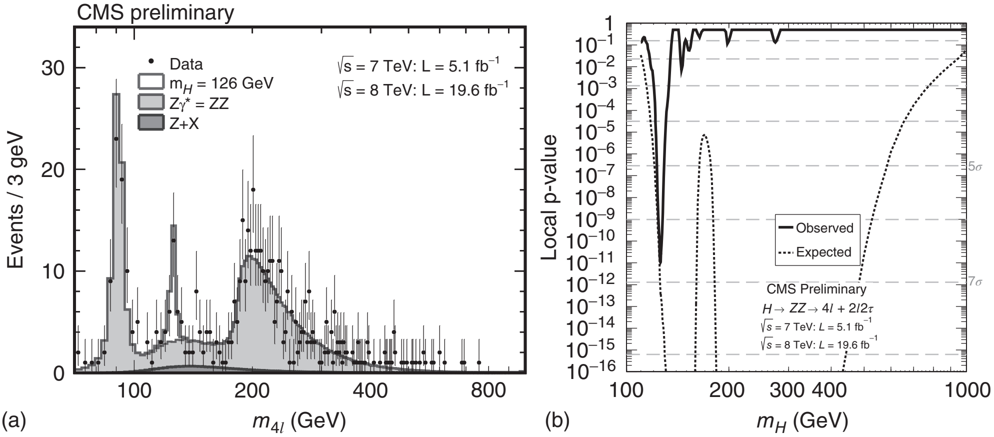


Figure 1.24 (a) CMS data. Distribution of the four-lepton reconstructed mass in the full mass range for the sum of the 4e, 4μ, and 2e2μ channels. Points represent the data, shaded histograms represent the background, and the unshaded histogram represents the signal expectation. The expected distributions are presented as stacked histograms. The measurements are presented for the sum of the data collected at $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV. No event is observed for $m_{4l} > 800$ GeV. (b) Significance of the local excess (right) with respect to the SM background expectation as a function of the Higgs boson mass in the full interpretation mass range 110–1000 GeV. (Reproduced with permission of [103].)

Table 1.5 shows a summary result of the spin-parity analysis.

The analysis (Table 1.5) favors 0^+ as the J^P assignment of the Higgs-like particle.

Coupling Strength The best way to determine whether the observed particle is the SM Higgs is to measure its coupling strength. In the SM, the coupling strength and the mass are directly related by $m_f = \gamma_f v/\sqrt{2}$ for the fermion and $m_V = g_V v/2$ for the vector boson where $v = 246$ GeV is the VEV of the Higgs. The probability to observe a particular Higgs production and decay mode, say, $gg \to h \to ff$, is proportional to

$$\sigma \cdot BR(gg \to h \to ff) = \frac{\sigma_{gg} \cdot \Gamma_{ff}}{\Gamma_H} = \sigma_{\text{SM}} \cdot BR_{\text{SM}} \cdot \left(\frac{\kappa_g^2 \kappa_{ff}^2}{\kappa_H^2} \right) \tag{1.116}$$

where Γ_H, Γ_{ff}, etc. are total decay width of the Higgs and partial decay width for $h \to ff$. The variables $\kappa_{ff}^2 = \Gamma_{ff}/\Gamma_{ff,\text{SM}}$ are the ratios of the coupling strength to those of the Standard Model. When the mass of the Higgs is fixed to the observed value $m_H = 125$ GeV, all the decay widths can be calculated. Therefore, the ratio $\mu \equiv (\sigma \cdot B)_{\text{ob}}/(\sigma \cdot B)_{\text{SM}}$ is a measure of how good the SM assumption is.

As an example, we show a test of the gluon-fusion production of the Higgs versus vector boson fusion production. Figure 1.26a shows the measured μs for various channels keeping only the two production couplings $\left((gg \to H) + (gg \to t\bar{t}H)\right.$ and the vector fusion process $\left.(V^* \to VH)\right)$ free. As one is looking at the same decay channels, all the uncertainties in the decay processes cancel each other and one

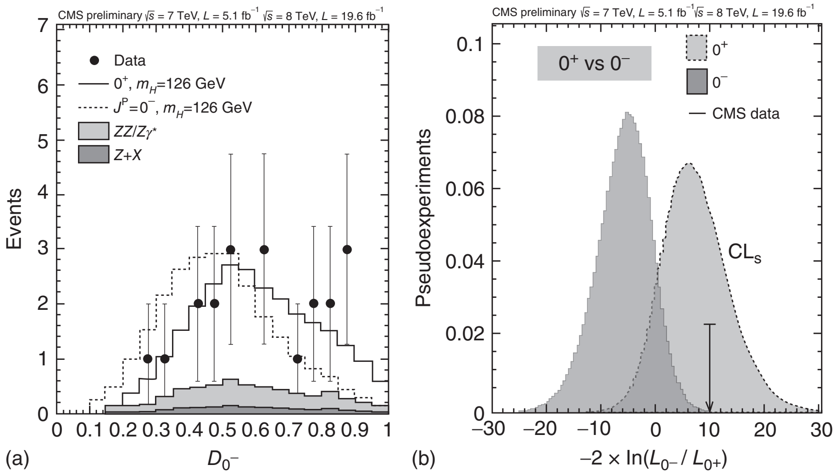


Figure 1.25 Spin analysis using $h \to ZZ^* \to 4l$ data, showing the kinematic discriminant built to describe the kinematics of production and decay of different J^P states of a “Higgs. The left figure shows $\mathcal{D}_{J^P}$ distribution with a requirement $\mathcal{D}_{bkg} = \mathcal{P}_{sig}/(\mathcal{P}_{sig} + \mathcal{P}_{bkg}) > 0.5$. The arrow in the right figure indicates the observed value. (Reproduced with permission of [103, 104].)

Table 1.5 List of models used in analysis of spin-parity hypotheses corresponding to the pure states of the type noted.(Reproduced with permission of [103, 104].)

J^P	Production	Commment	Expect $\mu = 1$	Observed 0^+	Observed J^P	CLs
0^-	$gg \to X$	Pseudoscalar	2.6σ (2.8σ)	0.5σ	3.3σ	0.16%
0^+_h	$gg \to X$	Higher dim operators	1.7σ (1.8σ)	0.0σ	1.7σ	8.1%
2^+_{mgg}	$gg \to X$	Minimal couplings	1.8σ (1.9σ)	0.8σ	2.7σ	1.5%
2^+_{mqq}	$gg \to X$	Minimal couplings	1.7σ (1.9σ)	1.8σ	4.0σ	<0.1%
1^-	$gg \to X$	Exotic vector	2.8σ (3.1σ)	1.4σ	$>4.0\sigma$	<0.1%
1^+	$gg \to X$	Exotic Pseudovector	2.3σ (2.6σ)	1.7σ	$>4.0\sigma$	<0.1%

The expected separation is quoted for two scenarios when the signal strength for each hypothesis is predetermined from the fit to data (0^P model) and when events are generated with SM expectation for the signal yield ($\mu = 1$). The observed separation quotes consistency of the observation with the 0^+ model or the J^P model, and corresponds to the scenario when the signal strength is predetermined from the fit to data. The last column quotes the confidence level criterion for the J^P model.

can see the difference of the production mechanism. No deviations from the SM is observed.

Because of multichannel productions in the hadron collider, the accuracy of the coupling constants that can be obtained at LHC has some limitations. We show how they can be improved if an electron–positron linear collider is used [Figure 1.26b] [106]. The discovered Higgs-like particle indeed behaves like the SM Higgs.

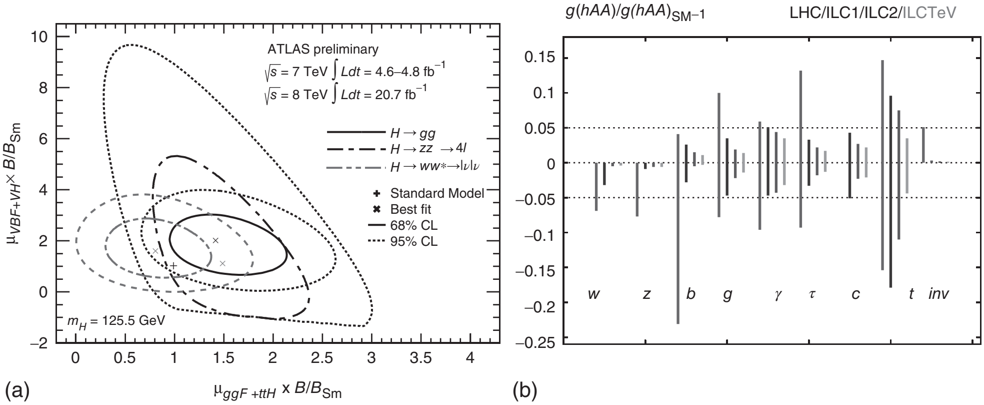


Figure 1.26 (a) Comparison of gluon–gluon fusion and vector boson fusion for the production of the Higgs and subsequent decays to various channels. $\mu = 1$ corresponds to the Standard Model. (Reproduced with permission of [105].) (b) Comparison of the capabilities of LHC and ILC for model-independent measurements of the Higgs boson couplings. The plot shows (from left to right in each set of error bars) 1σ confidence intervals for LHC at 14 TeV with 300 fb^{-1}, for ILC at 250 GeV and 250 fb^{-1} ('ILC1'), for the full ILC program up to 500 GeV with 500 fb^{-1} ('ILC2'), and for a program with 1000 fb^{-1} for an upgraded ILC at 1 TeV ('ILCTeV') (Reproduced with permission of [106].)

1.7.8
MSSM Higgs and Future Prospect

Neutral Higgs Next, we investigate whether the Higgs fits to the SUSY. The detection method for the neutral Higgs (H^0, h^0 A^0) in the MSSM at the electron collider is almost identical to that of the SM Higgs with increased degrees of freedom given by the existence of A^0. In addition to channels $e^-e^+ \to h^0 Z^*$, $h^0 A^0 \to \tau\tau b\overline{b}$, a four-jet final state was also available. Figure 1.27a shows the excluded region on the m_{h^0}-$\tan\beta$ plane obtained at the LEP collider. The discovered Higgs is also compatible with its interpretation as the light Higgs h^0 in SUSY for small ($\ll 1$) and large ($\gg 1$) $\tan\beta$.

Prospects at LHC The discovery method of MSSM Higgses at the hadron collider is similar to that of the SM Higgs. The difference is that the number of Higgs species has increased and it affects the decay branching ratios. Figure 1.27a shows the excluded region by LEP in the ($m_h - \tan\beta$) plane. Figure 1.27b shows the ATLAS team's work on their capability of finding one or more of the MSSM Higgs using various methods with an integrated luminosity of $300\,\text{fb}^{-1}$. Although parameters may have been taken a little optimistically, one sees that all the regions are covered. The LHC has the capability of discovering at least one of the neutral members of the Higgs in the MSSM [35, 94, 108].

If the light Higgs is discovered at LHC, one still has to distinguish whether it is the SM Higgs or that of SUSY. To pinpoint the parameter values, it is necessary to

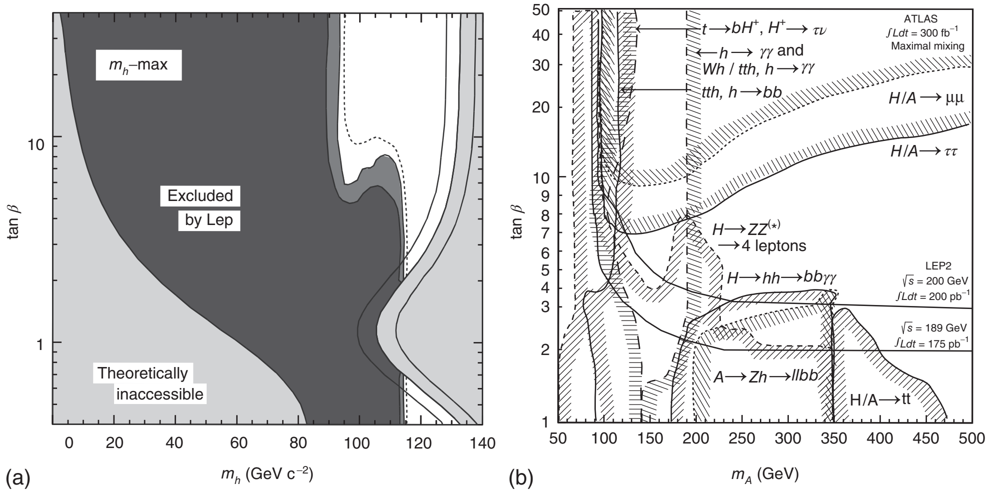


Figure 1.27 (a) MSSM light Higgs exclusion contours by LEP experiments, at 95% CL (light-green/middle-right thin gray region) and 99.7% CL (dark-green/dark gray region). The figure shows the excluded and theoretically inaccessible regions in the $(m_h, \tan\beta)$ projection. The upper edge of the theoretically allowed region is sensitive to the top quark mass; it is indicated, from left to right, for m_t = 169.3, 174.3, and 179.3. (Reproduced with permission of [7,107].) (b) Predicted 5σ discovery contours for MSSM Higgs boson detection in various channels in the $m_A - \tan\beta$ plane. (Reproduced with permission of [35,94,108].)

determine the decay branching ratios or to discover directly the charged Higgs or CP-odd A^0. Figure 1.28a shows the excluded regions by the LHC on the m_A-tan β plane prior to the discovery of the Higgs.

The discovery of the Higgs particles at $m_h = 125$ GeV, if interpreted as the light Higgs of the MSSM, constrains the allowed region. Figure 1.28b shows the allowed band (green band) at the tree level with $m_h = 125 \pm 3$ GeV limit [111]. The dark (blue) areas are the excluded regions by LEP and Tevatron/LHC [109, 110], and the gray area is the allowed region like the one shown in the left figure prior to the discovery of the Higgs particle. Note, however, that the limit is obtained by choosing other parameters governing the higher order corrections such that a maximum value for m_h is obtained ($m_h^{\max}$ benchmark scenario). While the lower limit is fairly stable, the upper limit is strongly dependent on the choice of the parameters. Therefore, at this stage, the green band should be interpreted only as the most favorable parameter region. The gray region above the band is disfavored but by no means prohibited.

Charged Higgs The charged Higgs can be produced at the LEP through the reaction

$$e^- + e^+ \rightarrow H^- + H^+ \tag{1.117}$$

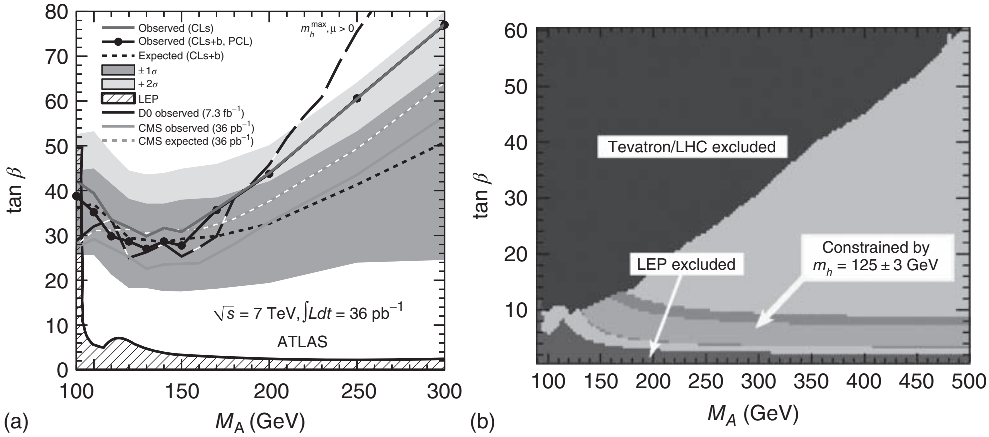


Figure 1.28 (a) Expected and observed exclusion limits in the $m_A - \tan\beta$ plane of the MSSM without the Higgs at 125 GeV. The region above the drawn limit curve is excluded at the 95% CL for the m_h = max scenario. The dark gray (green) and light gray (yellow) bands correspond to the $\pm 1\sigma$ and $+2\sigma$ error bands, respectively. The exclusion limits from LEP, D0, and CMS are also shown. (Reproduced with permission of [109, 110].) (b) Discovery of the Higgs particle at $m_h = 125$ GeV, if interpreted as the light Higgs of the MSSM constrains the allowed region on the $m_A - \tan\beta$ plane. The green horizontal band shows the region where m_h is compatible with the observed value within errors (125 ± 3 Gev). The brighter green is for the top mass value $m_t = 173.2$ GeV and the dark green corresponds to $\pm 1\sigma = 0.9$ GeV variation. The parameters governing the higher order corrections are taken such that a maximum value for m_h is obtained ($m_h^{\max}$ benchmark scenario). Dark (blue) areas are the excluded regions by LEP and Tevatron/LHC (see (a)). The gray area is the allowed parameter space prior to the latest LHC results. Note: Constraints on the upper edge of the green band is loose. This is still preliminary. If the m_h = max scenario is modified, the allowed region can be expanded to above significantly. (Reproduced with permission of [111–113].) (Please find a color version of this figure on the color plates.)

The Higgs decays preferentially to heavy particles. At the LEP, the main decay modes are $H^+H^- \to \tau^+\nu\tau^-\nu$ and $(c\bar{s})(\bar{c}s)$. Referring to Eq. (1.51), one can show

$$\frac{BR(H^+ \to \tau^+\nu)}{BR(H^+ \to c\bar{s})} \approx \frac{m_\tau^2 \tan^4\beta}{3m_c^2} \tag{1.118}$$

As the detection of the H^+ in both leptonic and hadronic decay modes is possible in the electron collider, the lower limit of the m_{H^+} independent of the value of $\tan\beta$ was obtained by the LEP2 [114].

$$m_{H^\pm} > 79.3\,\text{GeV} \quad [8] \tag{1.119}$$

At the hadron collider, the charged Higgs boson can be produced in different modes, but for $m_{H^\pm} < m_t - m_b$ it can be produced via the decay $t \to bH^+$, $H^\pm \to \tau\nu$. In the top quark decays, the W bosons decay equally to leptons of the three generations, while H^+ may decay predominantly into $\tau\nu$. Hence, an excess of $t\bar{t}$

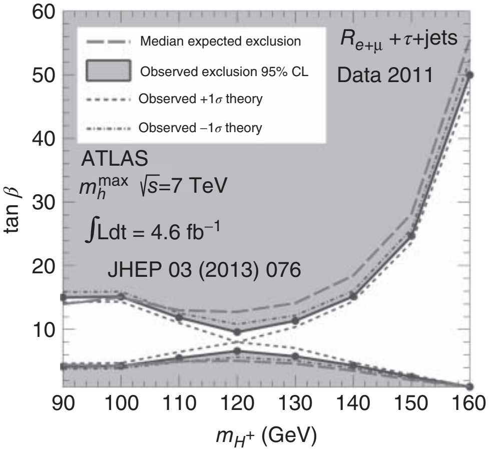


Figure 1.29 Summary of the 95% CL exclusions in the (m_{H^+}, $\tan\beta$) plane obtained by ATLAS [115]. The mass range below 90 GeV was excluded by the Tevatron. [116–118] (Reproduced with permission of [115].)

events with at least one hadronically decaying τ lepton (τ_{had}) in the final state, as compared to the rate for $t\bar{t}$ events with only electrons and/or muons, is a signature for charged Higgs bosons. One can use the ratio R_l to determine

$$R_l = \frac{BR(t\bar{t} \to b\bar{b} + l\tau_{\text{had}} + N\nu)}{BR(t\bar{t} \to b\bar{b} + ll' + N\nu)} \tag{1.120}$$

where ll' stands for electrons and muons with $l \neq l'$, and $N\nu$ stands for any number of neutrinos. Figure 1.29 shows the excluded region [115].

The mass region below 90 GeV was excluded by the Tevatron experiments [116–118].

As the mass of the charged Higgs is constrained by the relation Eq. (1.59), the discovery of the neutral Higgs at $m_h = 125$ GeV constrains the minimum mass value of the charged Higgs to $m_{H^\pm} > 152 \sim 161$ GeV [111]. This means that the detection of the charged Higgs in the decay channel of $t \to b + H^\pm$ is nearly excluded.

1.8
Summary

The discovery of the Higgs filled the last vacancy of the SM particles. But it has posed a new problem. The SM explains the EW data so well that it has pushed the boundaries of new physics beyond the limits predicted by many theoretical considerations. The mass of the detected Higgs is quite consistent with the predictions made by using the radiative corrections of the SM and the precision EW data obtained at LEP and Tevatron. The problem is that the Higgs is too light. The SM corrections to the Higgs mass diverge quadratically and, in order to reproduce the observed Higgs mass, the cutoff energy scale is constrained to be at most $\sim O(1\text{TeV})$. It contradicts the EW precision data analysis that no new physics should come in below $O(10)$ TeV. This is the little hierarchy problem.

One possible remedy was the SUSY. It was an aspired-for model because of its many desirable features. The mass value 125 GeV is within the range of predictions of SUSY. But it lies almost at the extreme limit of the allowed range. Many options of the SUSY were excluded. More details of the SUSY will be discussed in Chapter 5. The composite Higgs is another possibility. However, the composite models tend to produce a large mass, and the observed mass is at the low-end limit. To understand the dynamics that govern the Higgs, we need more experimental inputs.

2
Neutrino

2.1
Introduction

The flavor structure of the quarks and leptons is an unsolved problem in the Standard Model (SM). There are no hints to explain the three and only three repeated generations of the family consisting of a pair of quark and lepton doublets. The neutrino mass and flavor structure is probably the key to elucidate its mystery. The neutrino is unique among the elementary particles.

First of all, it has neither color nor an electric charge; hence it is the fundamental fermion that feels only the weak force in addition to the gravitational force. This fact has made it a unique tool for investigating the fundamental characteristics of weak interactions and probing the inner structure of hadrons by way of deep inelastic scattering.

Second, it is the only particle in the SM that lacks its right-handed partner.1) This fact was instrumental in formulating the SM. The fact manifests itself as the vector–axial vector (V–A) structure of the weak interaction, and pursuit of its origin culminated in the chiral gauge theories that lay the foundation of modern particle theories.

Third, the neutrino is the most abundant matter particle in the universe, with a billion neutrinos for each of the other known matter particles; only the ubiquitous photon is more abundant. Because of this, study of the neutrino has a great impact on cosmology.

Fourth, the neutrino masses are tiny compared to those of all other fundamental fermions. The seesaw mechanism, the prime scheme to explain the tiny mass, requires the presence of very heavy Majorana neutrinos, a vital clue to physics at a much higher energy scale. It may be also a clue to the matter–antimatter asymmetry of the universe.

Fifth, the neutrino itself could be a Majorana particle, a fermion that cannot be distinguished from its own antiparticle.

---

1) The statement is not quite true because the neutrino oscillation means finite mass, and hence the existence of the right-handed neutrino. However, the mass is so tiny that it is effectively true in most of the following discussion except the case when the mass itself is discussed.

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

Sixth, the neutrino oscillation provides, so far, the only experimental evidence for and a vital clue to the physics beyond the SM.

Last but not least, the neutrino intervenes in stellar evolution as well as in thermal cosmology in many important ways. The supernovae are neutrino bombs that distribute heavy elements in the universe. Planets that contain them are fertile and the genesis of all life forms. The hydrogen burning in the main sequence stars lasts for as long as 10 billion years because of the neutrino's participation and allows evolution of life forms including human beings. All these assets make the neutrino unique among the elementary particles and make its study an indispensable subject in understanding and probing particle physics.

2.2
Neutrino Mass

The neutrino mass problem is a very old theme, yet it has always been at the forefront of research fields. When Pauli predicted the existence of the neutrino for the first time in 1930, he already suspected its near-zero mass (by the standard of the time). Even today, direct measurements always end up with upper limits of the mass value. This fact and the discovery of the parity violation had led people at one time to suspect whether the neutrino is a two-component Weyl spinor (see Section 4.1.2 of [1]). Experimentally, only the left-handed neutrino ν_L and its charge conjugate ν_R^c[2] were observed. In conventional thinking, it should be the right-handed anti-neutrino. Experimentally, there are no evidences that the right-handed neutrino exists, but if it does, the proof of its existence is hard to come by. In the Standard Model, it is electrically neutral and constitutes a (weak) isospin singlet. Therefore, it does not interact weakly or electromagnetically. Hence, if it exists, there is no means to prove its existence.[3] Theoretically, however, there is a possibility that the neutrino is a Majorana particle, that is, a spin-1/2 particle which is its own anti-particle (see Section 4.4 of [1]).

On the other hand, if the neutrino is massive, as was verified by the discovery of the neutrino oscillation, ν_R has to exist. It follows from a simple argument. If massive, it can only fly at subluminal speed, and hence there always exists a Lorentz frame moving faster than the particle. In that frame, the neutrino is flying backward but the orientation of rotation, that is, its spin direction, does not change, and hence ν_L becomes ν_R in this frame. In the Standard Model, the neutrino is assumed to have zero mass. ν_R is sterile because it does not feel the electroweak force; hence there is no way to prove its existence, much less to investigate its properties. Hence, the assumption of its absence is the simplest and most economical solution to the problem. It does no harm in discussing most topics covered in this textbook.

Note, however, it is not impossible within the Standard Model to conceive the ν_R interaction. For instance, a Yukawa-type Lagrangian $\mathscr{L} \sim g\overline{\nu}_L \nu_R \phi$ is allowed in the framework of $SU(2) \times U(1)$ gauge symmetry. It is analogous to the mass-generating mechanism for the u quark. In this case, the vacuum expectation value (VEV) of ϕ gives mass to ν. That is, the existence of ν_R is necessary for the finite mass but not sufficient in the Standard Model. The coupling strength to the Higgs has to be finite, too. The ground for assuming the vanishing mass was the small experimental upper limit $m_\nu \lesssim O(1)$ eV. Since the difference of the mass is ascribed to the different coupling strength of the Higgs in the Standard Model, the tiny but finite mass of the neutrino translates to extremely small Higgs coupling to the neutrino. Naturally, one wonders why the neutrino has such a small coupling strength. The discovery of neutrino oscillation forced us to confront this problem.

---

[2] If the anti-neutrino exists as the charge conjugate of the neutrino, then $\nu_R^c = \overline{\nu}_R$. If there is no distinction between the neutrino and the anti-neutrino, $\nu_R^c = \nu_R$

[3] The neutrino can exert gravitational interaction. Therefore, it is possible that, if right-handed neutrinos exist in abundance, they could leave some trail from a cosmological point of view, but so far no observational evidences are available.

Let us review first what we know and what means are available to investigate the neutrino mass. Absolute values of the neutrino mass are not yet determined. Past direct measurements[4] have placed only upper limits on them [7].

$$\begin{array}{lcrl} m(\nu_e) & < & 2 & \text{eV} \\ m(\nu_\mu) & < & 190 & \text{keV} \\ m(\nu_\tau) & < & 18.2 & \text{MeV} \end{array} \tag{2.1}$$

Using cosmological arguments,[5] one can constrain the sum of the masses of all flavors to lie in the range (see Figure 2.1)

$$\textstyle\sum_j m_j \quad < \quad 0.14 \quad \text{eV} \tag{2.2}$$

A caveat is that the possibility of a heavy neutrino mass ($m_\nu > m_Z/2 = 45$ GeV) is not excluded.[6] We now have firm evidence that neutrinos are massive. From measurements of the neutrino oscillation (see Section 2.5),

$$\Delta m^2_{21} \simeq 7.7 \times 10^{-5}\ \text{eV}^2, \qquad |\Delta m^2_{31}| \simeq 2.4 \times 10^{-3}\ \text{eV}^2 \tag{2.4}$$

where only the difference of mass squared $\Delta m^2_{ij} = m^2_i - m^2_j$ is obtained. Remember that the oscillation experiments determine the difference of the mass eigenvalue squared, while the masses determined from direct measurements and given in Eq. (2.1) are in general weighted averages of the mass eigenvalues. If one member in the pair has a much smaller mass than the other, as is suggested by the seesaw mechanism $\left(m_i \sim m(l_i)^2/m_R,\ l_i = e, \mu, \tau,\ m_R \gg m_{l_i}\ \text{[see Eq. (2.19)]}\right)$, then

---

[4] The word "direct measurements" is used for experiments that can in principle determine the absolute value of each type of neutrino independently.

[5] Cosmological arguments include (i) neutrino decoupling (see arguments in Section 8.6.4), (ii) ΛCDM interpretation of CMB (cosmic microwave background) anisotropy (see Section 9.2), and (iii) large-scale structure of galaxy distribution, and BAO (baryon acoustic oscillation) [see arguments in Section 9.1.8, Eq. (9.88), and [119, 120, 121]].

[6] The large electron–positron (LEP) experiments on the invisible width only deny the existence of extra neutrinos that couple to Z and mass below $m_Z/2$. The cosmological limit obtained from the neutrino decoupling is obtained with the assumptions that the neutrinos are light ($\lesssim$1 MeV) and contain all types of neutrinos, that is, right-handed as well as sterile. If the cosmological expansion is interpreted with the assumption that the neutrino is much heavier than a few megaelectronvolts, the constrained range becomes ([125] (see arguments in Section 10.3.3)

$$\begin{array}{lcrl} m(\nu_j) & > & 8 & \text{GeV: Dirac neutrino} \\ m(\nu_j) & > & 14.4 & \text{GeV: Majorana neutrino} \end{array} \tag{2.3}$$

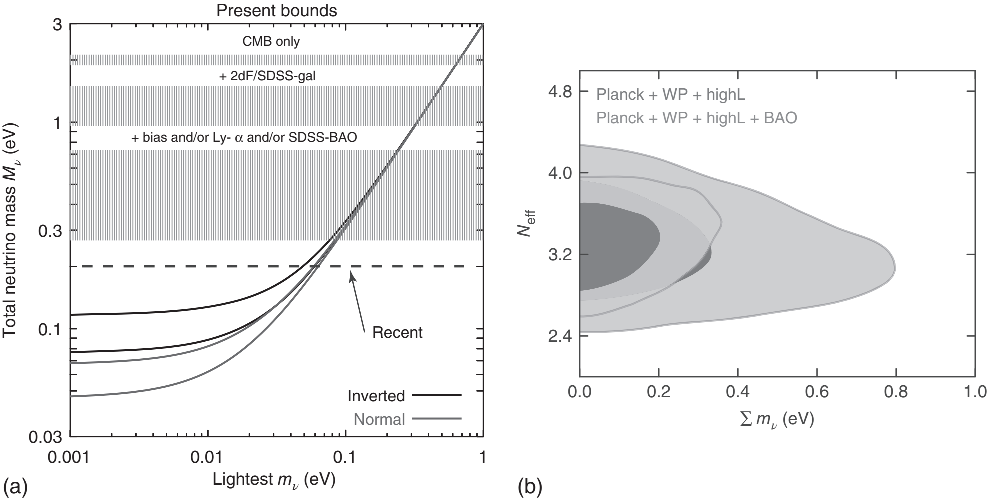


Figure 2.1 (a) Current upper bounds (95% CL) from cosmological data on the sum of neutrino masses compared to the values with neutrino oscillation data. For the meaning of lines, see Figure 2.45. Refer to Figure 2.41 for the distinction of inverted versus normal hierarchy. [119]. The dashed line is from more recent analysis [121, 122]. (b) Constraints on $\sum m_\nu$ versus N_{eff} from the CMB data. N_{eff} is the effective number of neutrinos at the time of neutrino decoupling. A finite value of $\Sigma m_\nu < 0.23$ eV for the flat universe is quoted [123, 124]. See discussions in Section 9.1.8 and footnote 19 in Section 8.6.5.

$|\Delta m^2_{ij}| \simeq \max[m_i^2, m_j^2]$. If this is true, $\max[m_\nu] \sim 0.05$ eV. One needs two orders of magnitude more accuracy for measurements of $m(\nu_e)$ to reach this value.

So far, the neutrino oscillation is the only experimental evidence that needs a new physics beyond the Standard Model. What role they play in the theory of elementary particle physics is a subject of forefront research. One problem that the neutrino has raised is that, if a theory can give a finite mass to it, it still has to explain why only the neutrinos have such a tiny mass compared to other leptons or quarks. The ratio $m_{\nu_e}/m_e \lesssim 10^{-5}$ is too small. If the origin of the mass is the coupling strength of the Higgs, there is no reason to treat the neutrino differently. Careful readers may have noticed that the top mass (~173 GeV) is larger than the electron mass (~0.5 MeV) by more than five orders of magnitude, and the mass hierarchy problem is universal and not restricted to the neutrino sector. However, the neutrino mass is special because it is extremely light within each generation ($m_{\nu_e} \ll m_e$, $m_{\nu_\mu} \ll m_\mu$, $m_{\nu_\tau} \ll m_\tau$). The heavy top mass is related to the riddle of the generation structure and should be considered as a problem in a different category. Having mentioned the generation puzzle, we note that there is no satisfactory explanation there either. Some consider the existence of a horizontal symmetry (i.e., across generations), but others consider that quarks and leptons are composites and the second and third generations are excited levels. But the mass hierarchy problem persists even in those models. String theories have models that ascribe it to the vacuum structure.

In the following, we start from general considerations of the mass term in the Lagrangian, including the possibility of the Majorana mass.

2.2.1
Mass Matrix

Majorana Mass For any two spinor fields ψ_1, ψ_2, the combination

$$\frac{m}{2}\left(\overline{\psi}_{1L}\psi_{2R}+\overline{\psi}_{1R}\psi_{2L}+h.c.\right) \qquad h.c. = \text{hermitian conjugate} \tag{2.5}$$

is Lorentz-invariant. When $\psi_1 = \psi_2$, the above term in the Lagrangian is referred to as the *Dirac mass term*. Consider a case where $\psi_1 = \psi_2^c$ or $\psi_2 = \psi_1^c$. It is still eligible to be called as the *self-energy of the field*. Then we have no reason to exclude it from the mass Lagrangian. The reason why it is not included usually in the mass term is because

$$\begin{aligned}\overline{\psi}^c_L\psi_R &= \left[\left(\frac{1-\gamma^5}{2}\right)\psi^c\right]^\dagger \gamma^0\psi_R = \psi_R^T C\psi_R \\ \overline{\psi}^c_R\psi_L &= \psi_L^T C\psi_L\end{aligned} \tag{2.6}$$

where $C(= i\gamma^2\gamma^0)$ is the charge conjugation matrix (see Appendix F of [1]). Equation (2.6) is not invariant under the phase transformation and, hence, does not satisfy the charge or lepton number conservation. It follows that the charged particle cannot have the mass term of this type; in other words, it cannot be a Majorana particle. What about the neutrino? In the Standard Model, it constitutes a doublet with the left-handed charged lepton. Therefore, it is natural to consider that the neutrino carries the same lepton number as the electron and, indeed, the weak interaction phenomenology is consistent with this notion. If it is true, the neutrino is not a Majorana particle either. But the neutrino being charge-neutral, there is no quantum number to differentiate it from its antiparticle experimentally. Besides, looking into the reactions hitherto considered as experimental evidences for the lepton number conservation

$$\begin{aligned}&\nu_\mu + n \to \mu^- + p, \qquad && \nu_\mu + p \nrightarrow \mu^+ + n \\ &\overline{\nu}_\mu + p \to \mu^+ + n, && \overline{\nu}_\mu + n \nrightarrow \mu^- + p,\end{aligned} \tag{2.7}$$

one realizes that the above reactions are nothing but a proof of the helicity conservation. This is because the V−A nature of the weak interaction restricts the lepton to be left-handed and the anti-lepton right-handed. That is, there exists no proof of the lepton number conservation for the neutrino. Therefore, if one accepts the possibility of the lepton number nonconservation, the most general Lagrangian for the free neutrino can be expressed as

$$\begin{aligned}-\mathscr{L}_\nu = &-\overline{\psi}i\gamma^\mu\partial_\mu\psi + m_D(\overline{\psi}_L\psi_R + h.c.) \\ &+ \frac{m_L}{2}\left(\overline{\psi}^c_R\psi_L + h.c.\right) + \frac{m_R}{2}\left(\overline{\psi}^c_L\psi_R + h.c.\right)\end{aligned} \tag{2.8}$$

The first term represents the kinetic term, and the rest are mass terms. The third and fourth terms are referred to as the *Majorana mass terms*. If they are absent, the above Lagrangian represents an ordinary Dirac particle. It is easier to grasp the picture if one introduces two Majorana fields defined by the following equations:

$$N_1 = \frac{1}{\sqrt{2}}\left(\psi_L + (\psi_L)^c\right), \qquad N_2 = \frac{1}{\sqrt{2}}\left(\psi_R + (\psi_R)^c\right) \tag{2.9}$$

N_1 and N_2 are referred to as left- and right-handed Majorana neutrino, respectively. These fields obviously satisfy the Majorana condition.[7] Conversely

$$\begin{aligned} N_{1L} &= \tfrac{1-\gamma^5}{2} N_1 = \tfrac{\psi_L}{\sqrt{2}}, \quad & N_{1R} &= \tfrac{1+\gamma^5}{2} N_1 = \tfrac{(\psi_L)^c}{\sqrt{2}} \\ N_{2L} &= \tfrac{(\psi_R)^c}{\sqrt{2}}, & N_{2R} &= \tfrac{\psi_R}{\sqrt{2}} \end{aligned} \tag{2.10}$$

Using the fact that any two fermion fields χ, ϕ satisfy the relations

$$\overline{\chi}_{L(R)}\gamma^\mu \phi_{R(L)} = 0, \qquad \overline{\chi}_{L(R)}\phi_{L(R)} = 0 \tag{2.11}$$

we can show that the following equalities hold:

$$\overline{\psi}\gamma^\mu\partial_\mu\psi = \overline{N}_1\gamma^\mu\partial_\mu N_1 + \overline{N}_2\gamma^\mu\partial_\mu N_2 \tag{2.12a}$$

$$\frac{1}{2}\left(\overline{\psi}^c_R\psi_L + h.c.\right) = \overline{N}_1 N_1 \tag{2.12b}$$

$$\frac{1}{2}\left(\overline{\psi}^c_L\psi_R + h.c.\right) = \overline{N}_2 N_2 \tag{2.12c}$$

$$\overline{\psi}_R\psi_L + h.c. = \overline{N}_1 N_2 + \overline{N}_2 N_1 \tag{2.12d}$$

Substituting Eq. (2.12) in Eq. (2.8), the neutrino Lagrangian is now expressed as

$$\begin{aligned} -\mathscr{L}_\nu &= -\overline{N}_1 i\gamma^\mu\partial_\mu N_1 - \overline{N}_2 i\gamma^\mu\partial_\mu N_2 \\ &\quad + m_D(\overline{N}_1 N_2 + \overline{N}_2 N_1) + m_L\overline{N}_1 N_1 + m_R\overline{N}_2 N_2 \\ &= -\overline{N}_1 i\gamma^\mu\partial_\mu N_1 - \overline{N}_2 i\gamma^\mu\partial_\mu N_2 + (\overline{N}_1, \overline{N}_2)\begin{bmatrix} m_L & m_D \\ m_D & m_R \end{bmatrix}\begin{bmatrix} N_1 \\ N_2 \end{bmatrix} \end{aligned} \tag{2.13}$$

where

$$\left[M\right] = \begin{bmatrix} m_L & m_D \\ m_D & m_R \end{bmatrix} \tag{2.14}$$

is called the *neutrino mass matrix*.

Inspecting the above Lagrangian, one sees that there are three cases consistent with experiments where the neutrino can have a finite mass.

(1) $m_D \neq 0$, $m_L = m_R = 0$: This is the Dirac neutrino which obeys the same equation of motion as the electron. One sees that the Dirac field is equivalent to two Majorana fields with identical mass. This provides an extra degree of freedom. One can combine the two Majorana fields to represent a complex spinor[8] whose Lagrangian satisfies the phase transformation invariance. In this way, the Dirac field can acquire an additional conserved lepton number.

The Dirac field contains both ν_L and ν_R with the same mass. In this case, absence of ν_R and vanishing mass are synonymous. This applies to any model without ν_R; for instance, the neutrino in the SU(5) GUT (grand unified theory) is massless (see Section 3.3 in Chapter 3). The mass-generating mechanism is identical to that of the electron, and its Feynman diagram is depicted in Figure 2.2a. As was stated earlier, one faces the hierarchy problem in this case.

---

[7] $N^c = \pm N$ and N satisfies the Dirac equation.

[8] The Majorana field may look complex, but it is superficial. One can choose a representation so that it is represented by a real field.

(2) $m_L \neq 0$, $m_R = m_D = 0$: It is a case one can acquire mass with only left-handed Majorana neutrino. If only the existence of ν_L is assumed as in the Standard Model, it can explain the finite mass of the neutrino in principle.

However, as ν_L is a member of a doublet having $I_3 = 1/2$, the mass term $\overline{(\nu_L)^c}\nu_L = -\nu_L^T C^{-1}\nu_L$ has $I_3 = 1$, and hence it constitutes a triplet. Since the Standard Model Higgs ϕ^0 has $I = 1/2$, it cannot couple to the neutrino at the tree level. One might consider the possibility of an extra new Higgs triplet $\chi = (\chi^0, \chi^-\chi^{--})$ which can acquire a VEV [126, 127]. The Lagrangian in this case takes the form

$$-\mathscr{L} \sim g(\overline{\nu}_L, \overline{e}_L)(\boldsymbol{\tau}\cdot\chi)\begin{bmatrix} e^c_R \\ -\nu^c_R \end{bmatrix} \tag{2.15}$$

This Lagrangian does not necessarily violate the lepton number conservation if χ has the lepton number 2, but even in this case it violates it after acquiring the VEV. As a result, two Goldstone bosons corresponding to the real and the imaginary part of χ^0 appear. They mix with neutral members of the Standard Model Higgs doublet, producing a massless Goldstone boson referred to as the *Majoron* and an additional light neutral scalar particle (denoted as ρ). They both couple to the Z boson and the neutrinos. The existence of the ρ (hence Majoron also) was denied by the LEP measurement of the invisible width of the Z. If χ does not carry the lepton number, the Majoron does not appear. But then, one has to assume the lepton-number-violating interaction and also has to make the Higgs coupling very small.

In order to generate the mass within the Standard Model, one needs an interaction of the type

$$-\mathscr{L} \sim \frac{f}{\Lambda}(\Phi^T C\boldsymbol{\tau}\Phi)\left\{(\nu_L, e_L)C\boldsymbol{\tau}\begin{bmatrix}\nu_L \\ e_L\end{bmatrix}\right\} \sim \frac{f}{\Lambda}\nu_L^T\nu_L\phi^0\phi^0 \tag{2.16}$$

as shown in Figure 2.2b. This Lagrangian is unrenormalizable, and therefore it has to be considered as a low-energy effective Lagrangian of a new physics (energy scale $\sim \Lambda$) producing radiative corrections. An example of such a new physics is depicted in Figure 2.2c [128, 129]. Here, one assumes the existence of two Higgs doublets and a charged scalar singlet h^- that violates the lepton number ($\mathscr{L} \sim -g\overline{l^c_R}\nu_L h^-$). One can construct a loop diagram which effectively produces the Lagrangian in Eq. (2.16). If ϕ^0 develops a VEV, the neutrino mass is given by $m_\nu \sim fv^2/\Lambda$. Assuming $f \sim 1$, $v = 246$ GeV, one needs $\Lambda \sim 10^{15}$ GeV to obtain $m_\nu \sim 0.1$ eV. If one asks unitarity for the

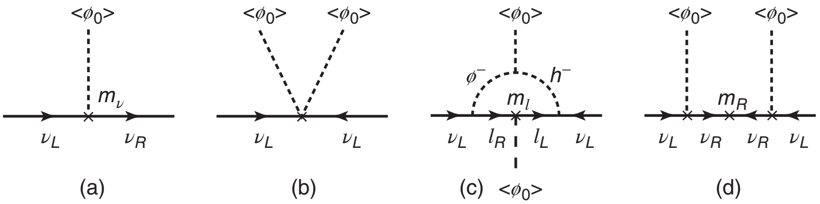


Figure 2.2 Various mechanisms to produce the neutrino mass. (a) Dirac neutrino. (b) Majorana neutrino. (c) An example of new physics that effectively leads to (b). (d) Seesaw mechanism. $<\phi_0>$ connected to × in the particle line means the mass term or interaction with hidden degenerate Higgs field, that is, vacuum.

reaction $\nu_L\phi \to \nu^c_R\phi$ all the way up to the Planck scale, one obtains $f/\Lambda < 2\pi/m_{\text{Planck}}$ and gets $m_\nu \lesssim 10^{-5}$ eV for $<\phi^0>= 246/\sqrt{2}$ GeV. If m_ν does not satisfy this condition, there must be a new physics scale smaller than the Planck energy [130, 131]. Thus, the small mass of the neutrino naturally leads to a new physics at the energy scale $\gtrsim 10^{15}$ GeV. GUTs are a natural framework for the neutrino.

(3) $m_D \neq 0$, $m_L \neq 0$, $m_R \neq 0$: This is the case where the right-handed as well as the left-handed Majorana neutrinos exist and which needs the full mass matrix in Eq. (2.14). It needs an extra ν_R with $m_R \neq m_L$ and is not compatible with the Standard Model. The interaction $\nu_L^T\nu_L\phi\phi$ can be effectively generated by exchanging the ν_R (see Figure 2.2d). In this case, $\Lambda = m_R$. The process is referred to as *the seesaw mechanism*.

Seesaw Mechanism Let us use ν', N to denote two fields with mass m_ν, m_N, which can be obtained by diagonalizing the mass matrix Eq. (2.14). The mass eigenstates are mixtures of the left- and right-handed neutrino and not in a definite chiral state. Let us assume that the Standard Model is correct except for the introduction of ν_R. Then, ν_L is massless and the Dirac mass is of the same order as that of charged leptons.

$$m_L = 0, \quad m_R \gg m_D \sim m_l \tag{2.17}$$

The mass eigenstates and eigenvalues are easily obtained.

$$\nu' \simeq N_1 - \frac{m_D}{m_R}, \quad N \simeq N_2 + \frac{m_D}{m_R} \tag{2.18a}$$

$$m_\nu \approx -\frac{m_D^2}{m_R}, \quad m_N \approx m_R \tag{2.18b}$$

The negative value of the mass eigenvalue is not important. By redefining $\nu \equiv \gamma^5\nu'$, one can make it positive. The above equation means

$$m_\nu \cdot m_R = m_D^2 \tag{2.19}$$

That is, by making m_R large, one can obtain a small m_ν, which is the origin of the name *seesaw mechanism*. The obtained state ν is almost left-handed and N is almost right-handed. The seesaw mechanism means the following relation:

$$\begin{aligned} m(\nu_e) : m(\nu_\mu) : m(\nu_\tau) &\sim m_e^2 : m_\mu^2 : m_\tau^2 \\ &= 2\times 10^{-10} : 10^{-5} : 4\times 10^{-3} \qquad \text{for } m_R = 10^{12}\ \text{GeV} \end{aligned} \tag{2.20}$$

where m_R is adjusted to reproduce the largest possible value of the neutrino mass obtained from oscillation measurements assuming $|m_i|^2 \gg |m_j|^2$. The magnitude of m_R suggests, again, that the new physics has energy scale close to that of GUT.

In summary, if the neutrino is a Dirac particle, one faces a theoretical problem that the coupling of the Higgs to the neutrino has to be set smaller by several orders of magnitude compared to that of the charged leptons, which is considered unnatural. But for the Majorana neutrino, several scenarios are possible to generate small masses. For this reason, the Majorana neutrino is more popular among theorists.

2.2.2
Left-Right Symmetric Model

Mass Hierarchy A prototype model in which the seesaw mechanism is naturally embedded is the left-right symmetric model [132, 133, 134]. The original motivation to introduce the left-right symmetric model was to consider that Nature respects perfect parity invariance at high energy and breaks it only at low energy spontaneously. It enlarges the $SU(2)_L$ in the Standard Model to $SU(2)_L \times SU(2)_R$ and assumes that the symmetry breakdown occurs in two steps:

$$SU(2)_R \times SU(2)_L \times U(1) \xrightarrow{<\Phi>_R} SU(2)_L \times U(1) \xrightarrow{<H>_L} U(1)_{EM} \tag{2.21}$$

$<\Phi>_R$ and $<H>_L$ mean that they develop VEVs to attach masses to right-handed and left-handed gauge particles, respectively. In the first stage, the gauge boson W_R which couples to right-handed particles obtains mass, and in the second step, the electroweak symmetry in the Standard Model is broken and W_L (=SM gauge boson) acquires mass. If W_R is heavy enough, the coupling of W_R to the fermion doublets is largely suppressed ($\sim m_{W_L}^2/m_{W_R}^2$) and the right-handed fermion doublets behave approximately as singlets of $SU(2)_L$. Therefore

$$m_R \simeq <\Phi>_R \gg m_D \simeq <H>_L \tag{2.22}$$

is a natural setting in the left- right symmetric model. The model appears naturally in the framework of the GUTs like $SO(10)$. Thus, enlarging the Standard Model to $SU(2)_R \times SU(2)_L \times U(1)$ is a possibility to explain the small mass of the neutrino.

Constraints on the Right-Handed Gauge Boson Phenomenologically, constraints for the left-right symmetric model can be given as follows: Since W_R and W_L couple to the same fermion doublets, they mix (mixing angle ζ) and constitute physically observed gauge bosons $W_1^\pm$, $W_2^\pm$, Z_1, Z_2. Under normal circumstances, the mixing angle is small. Then, $W_1 \approx W_L$, $W_2 \approx W_R$, $Z_1 \approx Z_L$, $Z_2 \approx Z_R$. If we assume $m_{W_R} \gg m_{W_L}$, consistency with the Standard Model is warranted. Numerically, one can obtain the following constraints:

(1) Limits on m_{Z_R}: If Z_R exists, one has to consider processes that exchange Z_R. This is an extra contribution to the neutral current (NC) processes, and hence changes the magnitude of the cross section calculated in the Standard Model. Since the SM reproduces data very well, one can set a lower limit on m_{Z_R} [135, 136]

$$m_{Z_R} > 1673\,\text{GeV} \quad (95\%\ \text{CL}) \tag{2.23}$$

(2) Limits on m_{W_R}:
Case: $m(\nu_R) \ll m_\mu$: In this case, ν_R appears as the decay product of μ. As the coupling is of the V + A type, one can detect the existence of W_R as a deviation from the standard V−A interaction. The most stringent restriction comes from measurements of the Michel parameters [137, 138, 139].

$$\begin{aligned} m_{W_R} &> 578\,\text{GeV}, \\ \zeta &< 0.02 \quad 90\%\ \text{CL} \end{aligned} \tag{2.24}$$

Case: $m(\nu_R) \gtrsim m_\mu$: If W_R exists, it contributes to the mass difference of the neutral K mesons through exactly the same process as W_L. If one assumes the left–right symmetric mixing matrix, its contribution can be calculated. The mass difference can be explained by the SM prediction alone. Then, using the standard assumption that the new physics contribution shall be less than the experimental value itself, one obtains [140, 141]

$$m_{W_R} > 2.5\,\text{TeV} \tag{2.25}$$

Direct Productions of Z_R, W_R If the accelerator energy is high enough, the new gauge bosons can be produced, and Z_R can be looked for in decays $Z_R \to l\bar{l}$. W_R can be looked for in decays $W_R \to l\nu_R$ if the mass of the right-handed neutrino is lighter than that of W_R. However, there are no constraints on $m(\nu_R)$ insofar as it does not couple to the Standard Model particles. Even if it is larger than $m(W_R)$, $W_R \to t_R b_R \to l + X$ is possible if $m(W_R)$ is larger than sum $m(t) + m(b)$.[9] Therefore, W_R can be searched in decays $W_R \to l + X$. Searches for Z_R, W_R at the Tevatron did not produce any positive signals, and the obtained upper limits were no better than those obtained from indirect arguments stated above. Searches for Z_R, W_R at the LHC gave

$$m(W') > 2.9\,\text{TeV} \quad 95\%\ \text{CL} \tag{2.26}$$

which will be treated in the next section when we discuss the possibility of new gauge bosons Z' and W' in general.

2.3
Electromagnetic Interaction

The neutrino is electrically neutral, and it does not interact electromagnetically at the tree level. However, it can do so radiatively. The electromagnetic properties of the neutrino can be obtained by investigating the following electromagnetic form factors (see Section 17.2 of [1]):

$$\begin{aligned}&<\nu(p', \lambda')|J^\mu_{EM}|\nu(p, \lambda)>\\ &\quad= \overline{u}(p', \lambda')\Big[\gamma^\mu \left\{F_1(q^2) - \gamma^5 G_1(q^2)\right\} + \sigma^{\mu\nu} q_\nu \left\{F_2(q^2) - \gamma^5 G_2(q^2)\right\}\Big] u(p, \lambda)\end{aligned} \tag{2.27}$$

where $q^\mu = p^\mu - p'^\mu$. The two terms in the first bracket $\{\cdots\}$ represent charge distribution, and the first term in the second bracket is the magnetic moment distribution. The last term represents the electric dipole distribution, which we will not discuss since it violates T invariance (see Section 9.2.2 of [1] and discussions on the electric dipole moment of the neutron).

Electric Charge Distribution For the left-handed neutrino, the γ^5 factor simply gives -1, so we write the form factor for the electric charge distribution as

---

[9] t_R can decay only to right handed particles. If $m(\nu_R) > m(W_R)$ the leptonic decay is forbidden. However, $t_R \to t_L \to l + X$ is possible because of large mass of t_R.

$Q(q^2) = F_1(q^2) + G_1(q^2)$. Denoting the charge distribution function as $\rho(\boldsymbol{r})$, where $\boldsymbol{r}$ is the position vector, its mean radius squared as $<r^2>$, and assuming isotropic distribution, the form factor in the static limit can be written as

$$Q(q^2) \equiv \int d\boldsymbol{r}\, \rho(\boldsymbol{r}) e^{i\boldsymbol{q}\cdot\boldsymbol{r}} = Q(0) - \frac{1}{6} q^2 <r^2> + \cdots \tag{2.28}$$

The first term is the total charge and vanishes, but the second term can be finite. Some processes that contribute to the second term are depicted in Figure 2.3.

The charge radius of the neutrino is not positive definite. Calculations give [142, 143, 131]

$$\begin{aligned}
<r^2>_{\nu_e} &\approx \left[+0.51 + 0.40\left\{\left(\frac{m_t}{175\,\text{GeV}}\right)^2 - 1\right\}\right] \times (10^{-16}\,\text{cm})^2 \\
<r^2>_{\nu_\mu} &\approx \left[+0.82 + 0.40\left\{\left(\frac{m_t}{175\,\text{GeV}}\right)^2 - 1\right\}\right] \times (10^{-16}\,\text{cm})^2 \\
<r^2>_{\nu_\tau} &\approx \left[+0.99 + 0.40\left\{\left(\frac{m_t}{175\,\text{GeV}}\right)^2 - 1\right\}\right] \times (10^{-16}\,\text{cm})^2
\end{aligned} \tag{2.29}$$

Experimentally, the charge radius can be determined from ν - e scattering. The effect appears as a deviation of the vector part of the Z^0 coupling constant $g_V \to g_V + \delta g_V$ or, equivalently, as $\delta \sin^2\theta_{WS}$.

$$\delta g_V = \frac{\sqrt{2}\pi\alpha}{3G_F} <r^2> = 2.38 \times 10^{30}\text{cm}^{-2} <r^2> \tag{2.30}$$

Setting the world average without ν_e scattering of $\sin^2\theta_{WS}$ (on shell) $= 0.231$ as the standard, one can determine the charge radius from ν_e scattering data [131]. Using the data by [144], one obtains

$$-5.3 \times 10^{-32} < \quad <r^2>_{\nu_\mu} \quad < 1.3 \times 10^{-32}\,\text{cm}^2 \tag{2.31}$$

If one uses the CHARM II data [145]

$$-0.77 \times 10^{-32} < \quad <r^2>_{\nu_\mu} \quad < 2.5 \times 10^{-32}\,\text{cm}^2 \tag{2.32}$$

For ν_e, one can use the Los Alamos Meson Physics Facility (LAMPF) data [146]

$$-5.0 \times 10^{-32} < \quad <r^2>_{\nu_e} \quad < 10.2 \times 10^{-32}\,\text{cm}^2 \tag{2.33}$$

The experimental limits are consistent with theoretical predictions that consider processes in Figure 2.3.

Magnetic Moment of the Neutrino From the properties of γ matrices, the following identity follows for any two fermion fields χ and ϕ:

$$\overline{\chi_L}\sigma^{\mu\nu}\phi_L = \overline{\chi_R}\sigma^{\mu\nu}\phi_R = 0 \tag{2.34}$$

This means that the magnetic interaction and the electric dipole interaction flip the helicity. Therefore, the two-component Weyl neutrino cannot have a magnetic moment. The four-component Dirac neutrino can flip its spin orientation, and

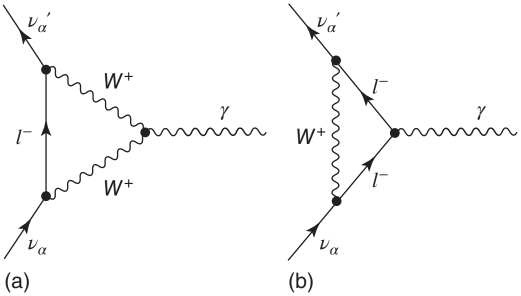


Figure 2.3 Diagrams that contribute to the neutrino's electromagnetic form factor.

hence it can have a magnetic moment. In order to investigate the case of the Majorana neutrino, we generalize Eq. (2.27) to include multiple flavors. Writing the Majorana neutrino as $N_i = n_{iL} + n^c_{i\,R}$, the magnetic interaction Lagrangian can be expressed as

$$-\mathscr{L} \sim \frac{\mu_{ij}}{2}\overline{N}_i\sigma^{\mu\nu}N_jF_{\mu\nu} = \frac{\mu_{ij}}{2}\left(\overline{n_{iL}}\sigma^{\mu\nu}n^c_{j\,R} + \overline{n^c_{i\,R}}\sigma^{\mu\nu}n_{jL}\right)F_{\mu\nu} \tag{2.35a}$$

Using

$$\begin{aligned} n^c_{j\,R} &= C\overline{n}^T_{jL}, \quad \overline{n}_{iL} = -n^{cT}_{iR}C^{-1}, \quad C^{-1}\sigma^{\mu\nu}C = -\sigma^{\mu\nu\,T} \\ \overline{n}_{iL}\sigma^{\mu\nu}n^c_{j\,R} &= -n^{c\,T}_{i\,R}C^{-1}\sigma^{\mu\nu}C\overline{n}^T_{jL} = -\overline{n}_{jL}\sigma^{\mu\nu}n^c_{i\,R} \end{aligned} \tag{2.35b}$$

we obtain

$$\mu_{ij} = -\mu_{ji} \tag{2.36}$$

Therefore, there exist no diagonal terms and the Majorana particle cannot have a magnetic moment by itself. However, if there are more than two Majorana fields, the interaction

$$n_i \to n_j + \gamma \tag{2.37}$$

is possible. That is, the Majorana particle can have a transition magnetic moment. If the neutrino has a magnetic moment, it means that it is a Dirac particle, but the existence of the magnetic interaction alone does not prove that the particle is of Dirac type.

If one calculates the magnetic moment using the loop diagrams in Figure 2.3, one gets a very small value [147, 148]

$$\begin{aligned} \mu_\nu &\equiv \kappa\mu_{\text{Bohr}} = \kappa\left(\frac{e}{2m_e}\right) \\ \kappa &\sim \frac{3G_F}{4\sqrt{2}\pi^2}m_\nu m_e \sim 3\times 10^{-19}\left(\frac{m_\nu}{1\,\text{eV}}\right) \end{aligned} \tag{2.38}$$

If the neutrino has a magnetic moment, its contribution increases the $\nu_e - e^-$ cross section by [149, 150]

$$\delta\left(\frac{d\sigma}{dE_e}\right) = \frac{\pi\alpha^2\mu_\nu^2}{m_e^2}\left[\frac{1}{E_e - m_e} - \frac{1}{E_\nu}\right] \tag{2.39}$$

where E_ν is the incident neutrino energy and E_e is the scattered electron energy.

The data from reactors and accelerators can be well reproduced with the Standard Model without the magnetic moment of the neutrino, which sets the limit on μ_ν.

$$\begin{array}{lcll} \mu_\nu(\nu_e) & < & 1.8\times 10^{-10}\mu_B & [151] \\ \mu_\nu(\nu_\mu) & < & 7.4\times 10^{-10}\mu_B & [146] \\ \mu_\nu(\nu_\tau) & < & 3.3\times 10^{-6}\mu_B & [152,\ 153] \end{array} \tag{2.40}$$

Stronger limits can be obtained by considering stellar cooling. In the stellar interior which is in the plasma state, the photon acquires mass (plasmon) and can decay into two neutrinos via e^-e^+ intermediate states (see Figure 2.55). As the neutrino interacts only weakly, it escapes from the star, carrying energy away with it. Therefore, the magnetic interaction provides additional stellar cooling, which is constrained by the standard stellar evolution model [154, 155], Another argument comes from the early universe evolution. When ν_L is in thermal equilibrium ($T > 1$ MeV) with electrons and photons, ν_R is produced as abundantly as ν_L, which increases the energy density, and hence the expansion rate. As a result, helium would be produced excessively [156, 157] (see arguments in Section 8.6.5). Combining all these arguments, one may take [131]

$$\mu(\nu_i) < (0.3\text{–}1)\times 10^{-11}\mu_B \tag{2.41}$$

A stronger limit is inferred from the observation of the supernova 1987A [158]

$$\mu(\nu_e) < (3\text{–}5)\times 10^{-13} \tag{2.42}$$

We will deal with this subject when we discuss SN1987A in Section 2.10.2.

In summary, both laboratory and astrophysical arguments give upper limits of the magnetic moment way above the theoretical limits. Speaking conversely, if one finds a value above the theoretical limit but below the mentioned phenomenological limits, it means a new physics.

2.4
Neutrino Mixing

Three flavors of the neutrino ν_f ($f = e, \mu, \tau$) are known but, in general, they do not necessarily coincide with the mass eigenstates ν_i ($i = 1 \sim 3$). In this case, the flavor eigenstates are expressed as mixtures of the mass eigenstates.

$$|\nu_\alpha \geq \sum_j U^*_{\alpha j}|\nu_j> \quad ^{10)} \tag{2.43}$$

When the mixing exists, the masses of the flavor eigenstates are the weighted means of the eigen-masses. Furthermore, if a certain flavor state appears, it also

---

10) We use notation defined by $U_{\alpha j} =<\nu_\alpha|\nu_j>$ where

$$|\nu_\alpha>= \sum_j |\nu_j><\nu_j|\nu_\alpha>= \sum_j <\nu_\alpha|\nu_j>^*\ |\nu_j> \tag{2.44}$$

The reason to use complex conjugate $U^*_{\alpha j}$ is to make it consistent with the PMNS (Pontecorvo–Maki–Nakagawa–Sakata) matrix defined by Eq. (2.107).

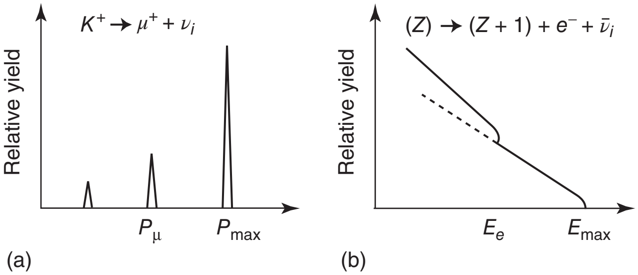


Figure 2.4 If neutrino mixing exists and the mass eigenstates coexist, several peaks/bumps appear in the spectrum.

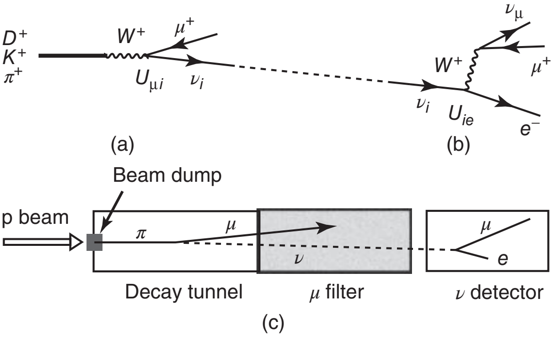


Figure 2.5 Feynman diagram for production (a) and decay (b) of a heavy neutrino. (c) Heavy neutrino detection using a beam dump and a neutrino detector.

means that all the coupled mass eigenstates are there. Consequently, we can expect the following phenomena [159, 160, 161]:

(1) Decays to the jth neutrino: $K^\pm(\pi^\pm \rightarrow \mu^\pm + \nu_j,\ e^\pm\nu_j)$ can happen with probability proportional to $|U_{lj}|^2$. The energy spectrum of the charged particles μ or e will have three peaks corresponding to the three masses m_j ($j = 1-3$), as illustrated in Figure 2.4a.

(2) In three-body decays like the beta decay, the energy spectrum will have step-like increase at the energy corresponding to $E_e = Q - m_j$ [$Q = M(Z) - M(Z+1)$], as illustrated in Figure 2.4b.

(3) If the neutrino is heavier than the electron or muon, it can further decay to these particles (Figure 2.5a,b).

A detection method for case (3) is illustrated in Figure 2.5c. Here, the beam–detector configuration is similar to that of a neutrino beam experiment (see Section 17.7 of [1]). The primary proton beam illuminates a thick material and produces many K's and π's. They decay in the decay tunnel, producing the heavy neutrinos which travel to the neutrino detector where they are detected. If one uses charmed particles ($D^\pm, D^0$) as the source of the heavy neutrinos, one does not need the decay tunnel because they decay promptly. Characteristic signals are pairs of leptons; especially, the μ–e pair constitutes an extraordinary signal. As the decay and detection probabilities are proportional to $|U_{li}|^2$ and $|U_{l'j}|^2$ respectively, one can measure $|U_{li}U_{l'j}|^2$ $(l, l' = \mu, e)$ [162].

(4) The neutrino oscillation [163, 164]: This is a phenomenon in which the flavor eigenstate changes as a function of time. For instance, if the ν_μ is produced at $t = 0$, it changes to ν_τ or ν_e as time passes. As it is the most important process in current neutrino researches, it will be described in detail in the next section.

Processes (1) and (3) use K and/or π, and hence they are suitable for searches of the neutrino in the mass range $1 \lesssim m_j \lesssim 490$ MeV. Process (2) uses nuclear beta decays, and hence it is suitable for searches in the mass range $1\,\text{keV} \lesssim m_j \lesssim 1$ MeV. So far, the massive neutrino in the above mass range has not been observed, and upper limits for the mixing matrix $|U_{lj}|^2 < 10^{-3} - 10^{-6}$ have been obtained (Figure 2.6 [165, 7]).

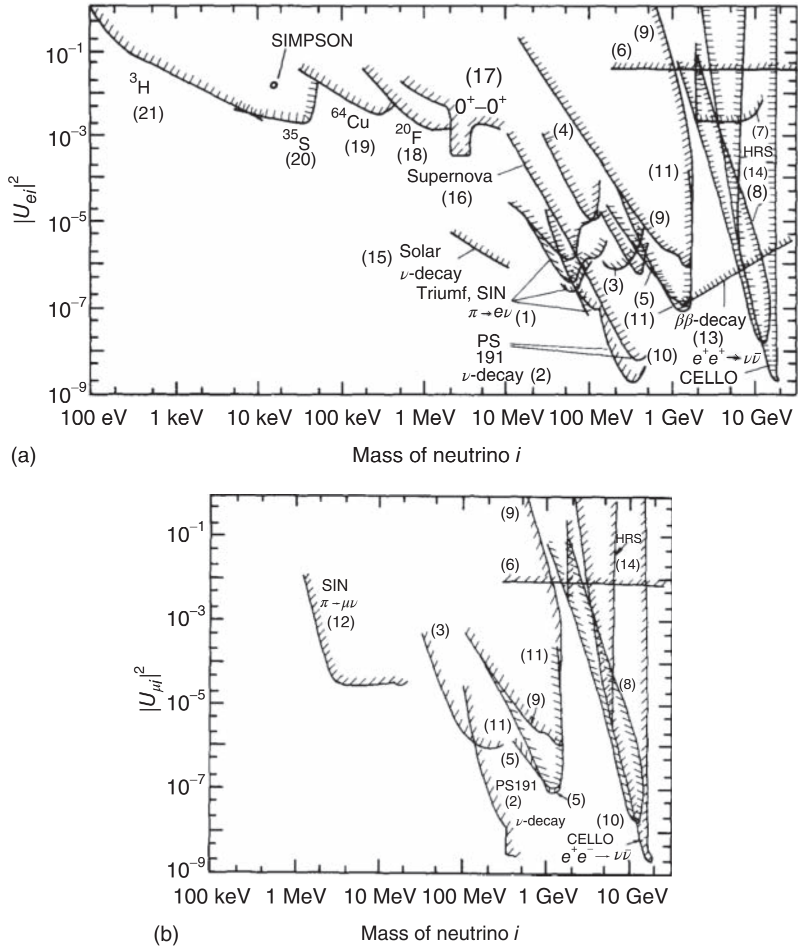


Figure 2.6 Heavy neutrino mixing upper limit as a function of the neutrino mass. (Reproduced with permission of [165].)

2.5
Neutrino Oscillation

Neutrino oscillation has been observed, which is evidence that the neutrino is massive. It is so far the only firm experimental phenomenon that goes beyond the Standard Model. By the neutrino oscillation, one cannot determine the absolute values of the neutrino masses but only measure the mass squared differences $\Delta m_{ij}^2 = m_i^2 - m_j^2$. However, if the neutrino mass is hierarchical, that is, $m_3 \gg m_2 \gg m_1$ as suggested by the seesaw mechanism, knowing the mass difference is essentially equivalent to knowing its absolute value. While the direct measurement of the neutrino mass in methods (1)–(3) is limited to $m \gtrsim 0.1$ eV with present technology, it has an advantage in that the oscillation experiments can explore an extremely small mass region, say, as small as $\Delta m^2 \simeq 1\ (\text{keV})^2 \sim 10^{-11}\ (\text{eV})^2$ (refer to Table 2.1). The oscillation experiments have found out the neutrino's finite mass and has the potential to determine all the mixing angles and the CP violating phase δ. Note, however, that the oscillation cannot determine the Majorana phases (see Eq. (2.109)). With present technology, they can only be determined by doing double-beta decay experiments, which will be described in Section 2.9.

2.5.1
Two-Flavor Oscillation

In order to treat the oscillation process precisely, one has to solve the equation of motion for transitions among the three flavors. However, many essential features of the oscillation can be studied using the two-flavor oscillation formula. Let us consider, for simplicity, the transition between ν_e and ν_μ. In this case, there is only one independent mixing parameter, which we choose to be the mixing angle θ. The flavor eigenstates can be expressed as follows:

$$\begin{aligned} |\nu_e> &= \quad \cos\theta|\nu_1> + \sin\theta|\nu_2> \\ |\nu_\mu> &= -\sin\theta|\nu_1> + \cos\theta|\nu_2> \end{aligned} \tag{2.45}$$

Expressing the neutrino energy as E_j, the mass eigenstates oscillate with time as

$$|\nu_j(t)>= |\nu_j(0)> e^{-iE_j t} \tag{2.46}$$

Table 2.1 Optimum facility for the neutrino oscillation.

ν Source	Energy (MeV)	Distance (m)	Δm^2 $(\text{eV})^2$
Accelerator	$10^3 - 10^5$	$10^2 - 10^3$ [a]	$10^{-3} - 100$
Pion factory	~ 30	$10 - 10^2$	$10^{-2} - 0.1$
Reactor	$1 \sim 2$	~ 20	$10^{-2} \sim 0.1$
Cosmic ray	$\sim 10^3$	$\sim 10^7$ [b]	$\sim 10^{-4}$
Sun	~ 1	$\sim 10^{11}$	$\sim 10^{-11}$
Supernova	~ 1	$50k$ ly	$\sim 10^{-20}$

[a] Short base line. Long baseline can be taken to be 100–10,000 km.

[b] Earth's diameter.

Therefore, the probability of a neutrino that was initially in the pure ν_e state to transform to the ν_μ state can be easily solved.

$$P(\nu_e \to \nu_\mu; t) = | <\nu_\mu(0)|\nu_e(t)> |^2 = \frac{1}{2}\sin^2 2\theta \left[1 - \cos(E_1 - E_2)t\right] \tag{2.47}$$

Using

$$E_2 - E_1 = \sqrt{p^2 + m_2^2} - \sqrt{p^2 + m_1^2} \simeq \frac{m_2^2 - m_1^2}{2p} \simeq \frac{m_2^2 - m_1^2}{2E} \equiv \frac{\Delta m_{21}^2}{2E} \tag{2.48}$$

we obtain the formula for the two-body oscillation probability. We will denote Δm_{21}^2 simply as Δm^2 in the following discussions.

$$\begin{aligned} P(\nu_e \to \nu_\mu; t) &= \sin^2 2\theta \sin^2\left(\frac{\Delta m^2}{4E} t\right) \\ &= \sin^2 2\theta \sin^2\left[1.27 \cdot \Delta m^2\,(\text{eV}^2) \cdot \frac{L\,(\text{km})}{E\,(\text{GeV})}\right] \end{aligned} \tag{2.49}$$

where $L = ct$ is the flight distance of the neutrino from production to detection. The oscillation wavelength is defined by

$$\lambda\,(\text{km}) = \frac{2\pi}{|E_1 - E_2|} \simeq \frac{4\pi E}{|\Delta m^2|} = 2.5\frac{E\,(\text{GeV})}{\Delta m^2\,(\text{eV})^2} \tag{2.50}$$

The probability for ν_e to remain as ν_e is given by

$$P(\nu_e \to \nu_e) = 1 - P(\nu_e \to \nu_\mu) \tag{2.51}$$

As can be seen from Eq. (2.49), the neutrino oscillation occurs only when there are both mixing ($\theta \neq 0$) and mass difference ($\Delta m \neq 0$). A measure of the observed mass difference is given by

$$\frac{L}{\lambda} \sim \Delta m^2 \cdot \frac{L}{E} \sim 1 \tag{2.52}$$

Therefore, by a suitable choice of E/L, one can measure a wide range of Δm^2's. Table 2.1 shows typical neutrino sources and measure of Δm^2.

Experimentally, there are two methods to detect the oscillation process: appearance and disappearance. In the former, one observes the appearance of ν_e in the downstream of a ν_μ beam, the initial condition being prepared by the accelerator. In the latter, one measures the neutrino flux at two points and detects whether a fraction of the initial flux disappears. Ideally, one wants to measure the flux as a function of the distance and draw a curve like the one illustrated in Figure 2.7a, but as the neutrino interacts only weakly, it is hard to construct an apparatus long enough to measure the flux variation. Technically, the appearance experiment is easier, but has poor control of the experimental conditions. The disappearance experiment, on the other hand, can measure a variety of parameters including the magnitude of the mixing matrix, flavor change like $\nu_f \to \nu_{f'}, \overline{\nu}_f, \nu_X$ (ν_X is any unknown neutrino), and so on. If one detects the oscillation, one can determine both the mixing ratio $\sin^2 2\theta$ and the difference of mass squared Δm^2. When only an upper limit is obtained, one usually describes the result as a contour of the

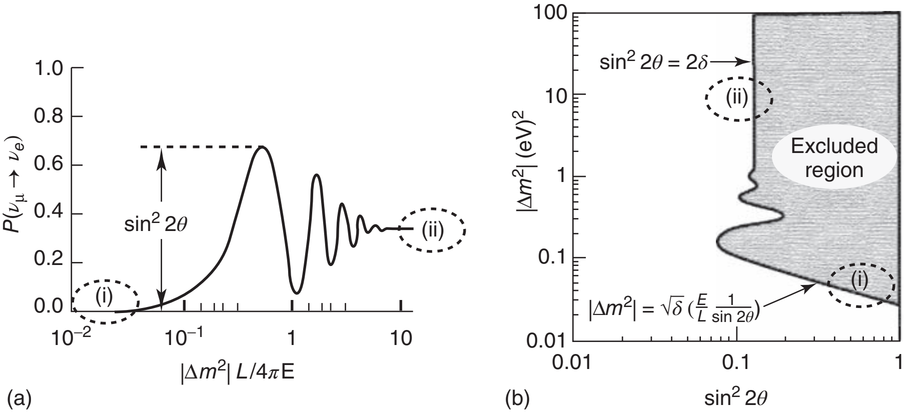


Figure 2.7 (a) Neutrino oscillation intensity plotted as a function of L/E. (b) Excluded region by a null experiment. The boundary is determined by the experimental error δ, but its shape reflects the oscillation amplitude. Region (i) is the long-wavelength limit and region (ii) is the long-distance limit.

excluded region on the $\sin^2 2\theta - \Delta m^2$ plane (Figure 2.7b). In the extreme case of the long-wave approximation ($L \ll \lambda$) where Δm^2 is small, Eq. (2.49) becomes

$$P(\nu_e \to \nu_\mu) \xrightarrow{\Delta m^2 \ll \frac{L}{E}} \sin^2 2\theta \left(1.27 \cdot \Delta m^2 \cdot \frac{L}{E}\right)^2 \tag{2.53}$$

This is illustrated as region (i) in Figure 2.7. When the oscillation is not observed, one can only determine the upper limit of the combined value

$$\sin 2\theta \cdot \Delta m^2 < \frac{E}{L}\sqrt{\delta} \tag{2.54}$$

where δ is the experimental error.

In the short-wave approximation ($L \gg \lambda$) where the adopted distance is too long compared to the oscillation wavelength (this is the region denoted as (ii) in Figure 2.7), the oscillation is violent and its intensity is averaged. The smearing occurs as a result of the finite energy width and spread of the production points. In this case, only the upper limit of the mixing parameter can be determined.

$$P(\nu_e \to \nu_\mu) = \sin^2 2\theta \ < \sin^2\left(1.27 \cdot \Delta m^2 \cdot \frac{L}{E}\right) > \simeq \frac{1}{2}\sin^2 2\theta < \delta \tag{2.55}$$

In accelerator experiments, $\delta \approx O\ (10^{-3} \sim 10^{-4})$ is possible but in astrophysical observations, sensitivity to the mixing angle is not so good ($\delta \gtrsim 0.1$). In order to investigate the small Δm^2 region, it is advantageous to have small E/L, that is, low energy and long distance. Referring to Table 2.1, one sees that the best sensitivity for the small Δm^2 is obtained from solar observations.[11]

Historically, all the early ($\lesssim$1990) accelerator and reactor experiments failed to detect the oscillation. This is because the observed value of Δm^2 was so small that a long distance (of the order of a few to several hundreds of kilometers) was required to make E/L to match Δm^2. It was natural that the first observation of the oscillation was made by using neutrinos of cosmic ray and solar origin. Only after knowing the exact parameter values could one design accelerator and reactor experiments, which we will describe later in Section 2.5.3.

---

[11] If one can detect the oscillation of the supernova neutrino, then Δm^2 is even smaller.

2.5.2
Atmospheric Neutrino

The first indication of the neutrino oscillation was a deficit of the ν_μ component compared to ν_e components in the observation, around 1990, of the atmospheric neutrino [166, 167, 168].[12] The atmospheric neutrino is produced when the primary cosmic rays (mainly high-energy protons floating in the galaxy) reach the Earth, interact with the atmosphere, generate showers by cascade interactions, and produce kaons and pions which decay to ν_μ and ν_e. As the muon flux at the Earth's surface can be measured, one can calculate back the primary cosmic ray fluxes, which, in turn, can be used to calculate the neutrino flux [170]. However, the mechanism of multiparticle production is not yet completely understood and differences among various calculations can be as large as 30% [171]. Therefore, obtaining the absolute cross section is difficult theoretically as well as experimentally. However, by taking ratios of the observables, major errors compensate each other and the relative error can be controlled to below 10%.

A major production–decay mechanism is given by

$$\begin{aligned} &p + A \to \pi^\pm\,(K^\pm) + X \\ &\qquad \pi^\pm\,(K^\pm) \to \mu^\pm + \overset{\subset\!-\!\supset}{\nu}_\mu, \qquad \mu^\pm \to e^\pm + \overset{\subset\!-\!\supset}{\nu}_\mu + \overset{\subset\!-\!\supset}{\nu}_e \end{aligned} \tag{2.56}$$

where $\overset{\subset\!-\!\supset}{\nu}$ denotes the neutrino or anti-neutrino depending on the charged partner. Naively, Eq. (2.56) would give a factor 2 for $N(\nu_\mu)/N(\nu_e)$, which is approximately right for low-energy ($<2-3$ GeV) neutrinos [171]. The Kamiokande group observed [166] that the measured ν_e flux reproduces theoretical predictions approximately, but that of the ν_μ was less than predictions. They presented their data as the double ratio.[13]

$$\begin{aligned} R &= \frac{(\nu_\mu + \bar{\nu}_\mu)/(\nu_e + \bar{\nu}_e)|_{\text{data}}}{(\nu_\mu + \bar{\nu}_\mu)/(\nu_e + \bar{\nu}_e)|_{\text{MC}}} \\ &= 0.658 \pm 0.016\,(\text{stat}) \pm 0.035\,(\text{syst}) \quad \text{sub-GeV} \\ &= 0.702 \quad {}^{+0.032}_{-0.030}\,(\text{stat}) \pm 0.101\,(\text{syst}) \quad \text{multi-GeV} \end{aligned} \tag{2.57}$$

where the subscript MC means Monte Carlo calculations based on theoretical models [170, 172]. Sub (multi)-GeV denotes visible energy in the detector $E_{\text{vis}} \lessgtr 1.33$ GeV. Another piece of evidence was provided by the IMB (Irvine-Michigan-Brookhaven) group [168]. An interpretation for the possible existence of the neutrino oscillation has been made. But because of the theoretical uncertainty on the flux as well as the existence of conflicting data [173, 174] which claim no anomaly, the hypothesis of the neutrino oscillation was not universally accepted.

---

[12] The solar puzzle that existed from the early 1970s turned out to be the earliest indication when it was proved to be an oscillation effect in 2002 [169].

[13] Their original values have been updated [171].

A definitive evidence for the oscillation was presented by the second-generation detector SuperKamiokande (hereafter abbreviated as SuperK) which measured the zenith angle distribution of the neutrino and showed up–down asymmetry [175, 176, 177]. The neutrino production rate at the Earth's surface is considered uniform everywhere except at very low energy. The low-energy primary cosmic rays below 2 GeV are captured by the Earth's magnetic field and tend to concentrate on the Earth's magnetic poles. The effect on the neutrino appears below $\lesssim$500 MeV, which can be neglected or corrected. The neutrinos with a zenith angle greater than 90° are produced at the rear side of the Earth and reach the detector with a flight distance determined by the zenith angle (Figure 2.8a). In other words, one is measuring the neutrino intensity as a function of the flight distance, an ideal setting for observing the neutrino oscillation. As the Earth is transparent to neutrinos, the flux within a fixed solid angle from above and below should be the same in the approximation of neglecting the thickness of the atmosphere (Gauss's theorem in electromagnetism). Therefore, the zenith angle distribution should be flat in this approximation. The effect of the finite thickness enhances the flux in the horizontal direction (i.e., at $\theta = \pi/2$). This is because the flight length of the primary cosmic ray becomes larger, producing more neutrinos. However, the up–down asymmetry should not be affected.

The zenith distributions of the atmospheric neutrino are shown in Figure 2.8b.

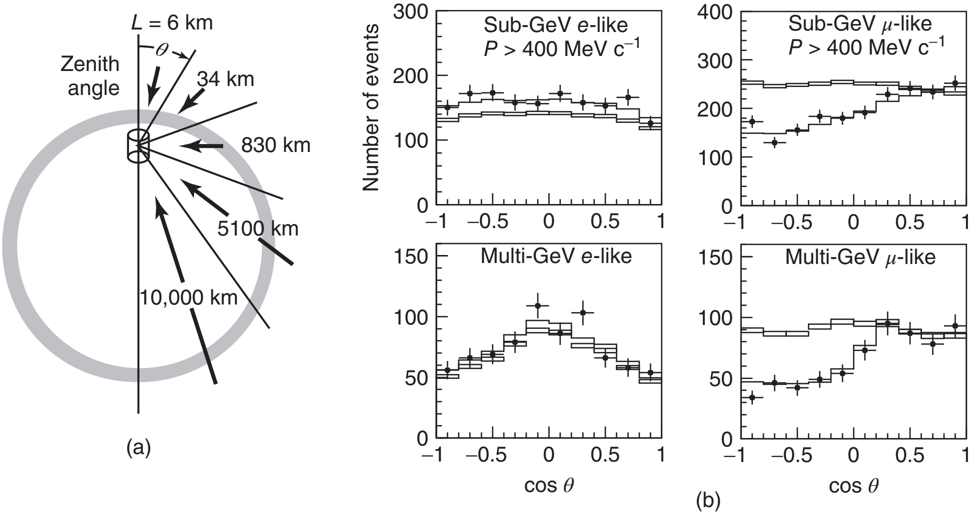


Figure 2.8 (a) For atmospheric neutrinos, zenith angle θ is directly related to the flight distance L. (b) Zenith angle distributions of the atmospheric neutrinos for sub-GeV data (top) and multi-GeV (bottom) data. The zenith angle distributions for fully contained one-ring e-like and μ-like events with visible energy < 1.33 GeV (sub-GeV) and >1.33 GeV (multi-GeV) are shown. The boxed histograms show the Monte Carlo events with no oscillation, and the solid histograms show the best-fit expectations for $\mu_\mu \rightarrow \nu_\tau$ oscillations. (Reproduced with permission of [7].)

One sees that the boxed histograms, which express the yields of MC simulation without the oscillation, are flat for the sub-GeV (visible energy <1.33 GeV) but peak at the center for multi-GeV (>1.33 GeV). The observed data for the electron shows no or little deviation from the prediction, but the muon data shows a clear anisotropy, confirming the previously observed muon deficit. As the Earth is transparent to neutrinos, the origin of the asymmetry can be attributed only to the difference in the flight path length. Therefore, it is hard to think of other reasons than neutrino oscillations. The data agrees well with the hypothesis of the neutrino oscillation. As no significant reduction is observed for ν_e, one can assume that ν_e did not transfer to ν_μ which also means no inverse process $\nu_\mu \not\to \nu_e$ by T invariance. Reactor experiments [178] at that time did not observe $\overline{\nu}_e \to \overline{\nu}_x\ (x = \mu, \tau)$, either. By conjugation–parity transformation–time reversal (CPT) invariance, it also means $\nu_e \not\to \nu_x$. Therefore, one can infer that the disappearance of ν_μ means $\nu_\mu \to \nu_\tau$ oscillation.[14] The solid histograms in Figure 2.8b are obtained using this hypothesis with parameters

$$\sin^2 2\theta > 0.92, \qquad 1.5 \times 10^{-3} < |\Delta m^2| < 3.4 \times 10^{-3} \tag{2.58}$$

Figure 2.9 shows the same data not as a function of the zenith angle but as a function of L/E, which is easier to understand the oscillation phenomenon intuitively.

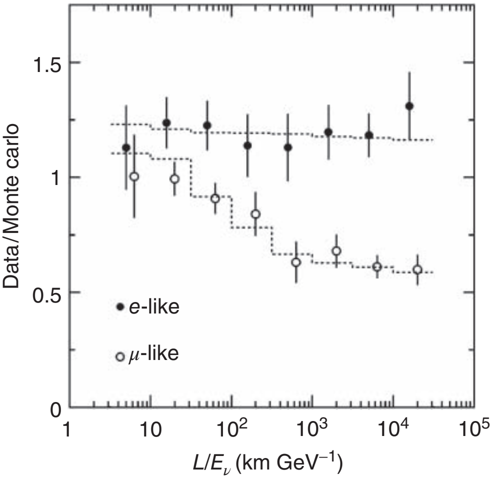


Figure 2.9 Ratio of the number of FC (fully contained) data events to FC Monte Carlo events versus reconstructed L/E_ν. The points show the ratio of observed data to MC expectation in the absence of oscillations. The dashed lines show the expected shape for $\nu_\mu \to \nu_\tau$ at $\Delta m^2 = 2.2 \times 10^{-3}$ eV^2 and $\sin^2 2\theta = 1$. The slight L/E_ν dependence for e-like events is due to contamination (2.7%) of ν_μ CC (charged current) interactions. (Reproduced with permission of [175].)

---

[14] With the given data here, one can also interpret the process as $\nu_\mu \to \nu_x$, where ν_x is the sterile neutrino that does not couple to the SM gauge bosons. Later analyses including other flavors have confirmed the $\nu_\mu \to \nu_\tau$ oscillation possibly with a small mixture of $\nu_\mu \to \nu_e$

2.5.3
Accelerator Experiments

Inspired by the observations and knowing the parameters of the atmospheric and later of the solar neutrino oscillation, long-baseline ($L \gtrsim 100\,\text{km}$) accelerator/reactor experiments were carried out. Figure 2.10 shows three major sites of the world's long-baseline oscillation experiments. The basic concepts of the neutrino beam and the neutrino detector were described in Section 17.7 of [1]. Here, we will limit our description of experimental apparatus to those specific to the oscillation process.

For experiments where two detectors are placed at a distance of over 100 km apart, GPS (global positioning system) is used to find the accurate position of the far detector and direct the neutrino beam to a pinpoint target and to synchronize timing with the near detector (Figure 2.11). Figure 2.12 illustrates an example of synchronizing the far detector events with the accelerator timing.

K2K (KEK to Kamioka) [180, 179] is the first successful accelerator-based experiment. The group extracted the 12-GeV proton beam from KEK/PS (proton

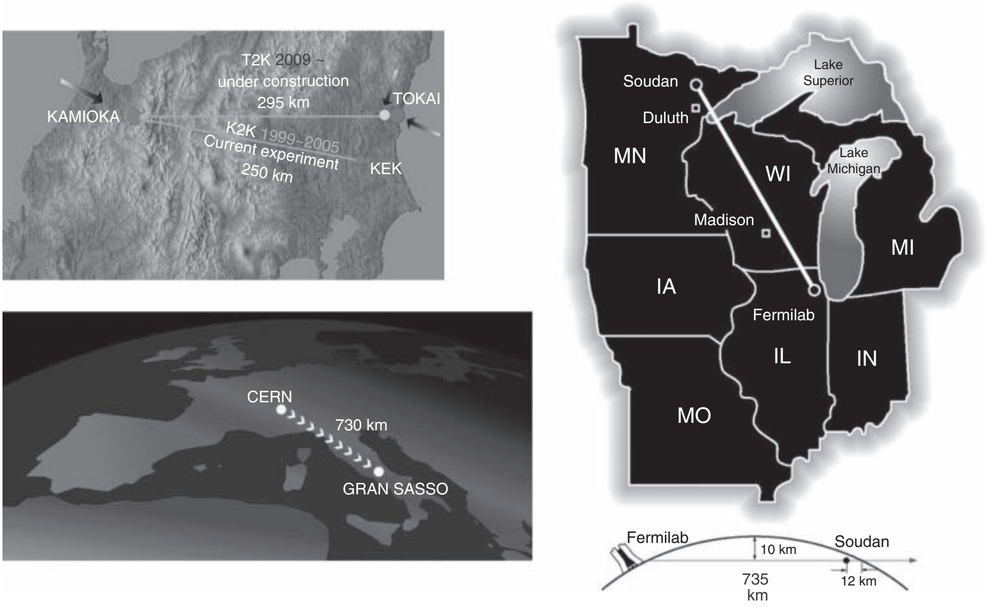

T2K 2009 ~
under construction
295 km
TOKAI
KAMIOKA
K2K 1999~2005
Current experiment
250 km
KEK
CERN
730 km
GRAN SASSO
Soudan
Duluth
Lake
Superior
MN
WI
Lake
Michigan
Madison
MI
IA
Fermilab
IL
IN
MO
Fermilab
10 km
Soudan
735
km
12 km

Figure 2.10 World neutrino oscillation experiment sites. From top clockwise: K2K (KEK to Kamioka, 250 km), T2K (Tokai to Kamioka, 295 km) in Japan, MINOS (Main Injector Neutrino Oscillation Search, Fermilab to Soudan, 735 km) in the United States, and CNGS (CERN to Gran Sasso in Italy, 730 km) in western Europe. (Please find a color version of this figure on the color plates.)

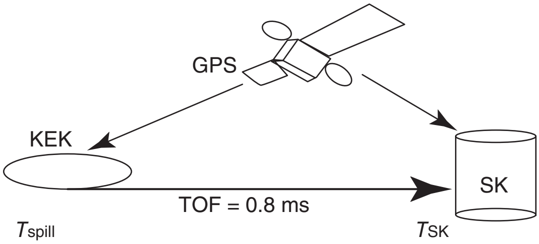


Figure 2.11 For the long-baseline experiment, GPS (global positioning system) is used to find accurate position of the far detector and synchronize timing of the near and far detectors.

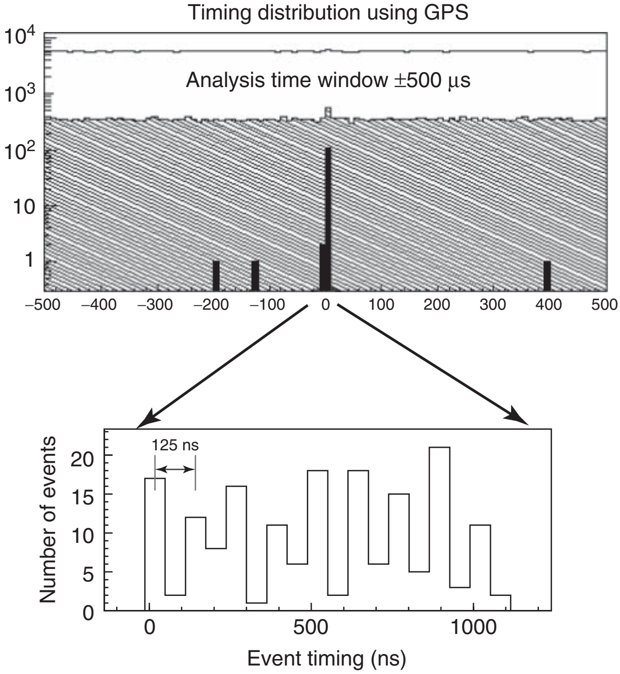


Figure 2.12 Time distribution of events at the K2K far detector. Background events are distributed uniformly as a function of time (top). Signal events are shown as a thick line histogram. Events of accelerator origin are in a time window −0.2 to + 1.3 μs. Their time distribution reveals a characteristic accelerator bunch structure. (Reproduced with permission of [179].)

synchrotron) and generated a low-energy ν_μ beam at the right value of $L/E_\nu \sim 520\,(L = 250\,\text{km}, <E_\nu> = 1.3\ \text{GeV})$.[15]

---

[15] $X_{32} \equiv \Delta m^2_{\text{atm}} L/4E = \pi/2$ is the oscillation maximum or dip in the transition probability, which corresponds to $E_\nu \simeq 0.48$ GeV [see Eq. (2.59) and Figure 2.13]. At higher energies, the transition probability goes up but the neutrino beam flux goes down, producing a peak around $E_\nu \sim 1.3$ GeV.

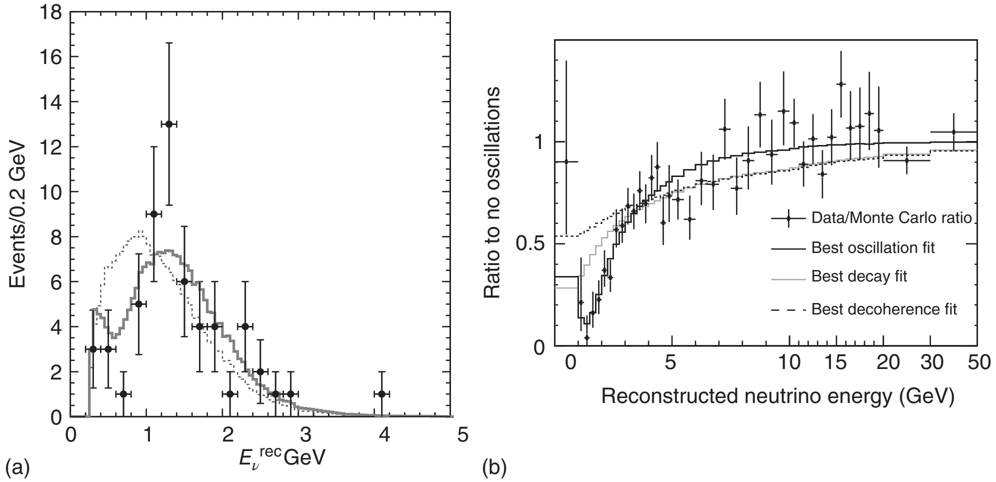


Figure 2.13 (a) K2K oscillation data. The reconstructed E_ν distribution for the one-ring μ-like sample. Points with error bars are the data. The solid line is the best-fit spectrum with neutrino oscillation, and the dashed line is the expectation without oscillation. These histograms are normalized by the number of observed events. (Reproduced with permission of [179].) (b) Ratio of the MINOS far detector data and the expected spectrum for no oscillations. The best-fit with the hypothesis of $\nu_\mu \to \nu_\tau$ oscillations as well as the best fit to alternative models (neutrino decay and decoherence) are also shown. (Reproduced with permission of [8, 181].)

The first (near) detector was placed on KEK site, and the SuperK water Cherenkov detector 250 km away was used as the second (far) detector to catch the ν_e component in the ν_μ beam. The total number of protons on target (POT) for physics analysis amounted to 0.92×10^{20}. The observed number of beam-originated, fully contained events in the 22.5-kt fiducial volume of SuperK was 112, to be compared with expected number of $158^{+9.2}_{-8.6}$ events without oscillation. Figure 2.13a shows the data as a function of the neutrino energy. The expected distributions with and without oscillation are also shown. A dip appears at the position where it should be, and confirms the atmospheric neutrino oscillation data obtained by the SuperK.

MINOS: The second long-baseline accelerator experiment Main Injector Neutrino Oscillation Search (MINOS) [182] used the NUMI (Neutrinos at the Main Injector) beam, which is a Fermilab facility to use the primary proton beam with energy 120 GeV. The far detector is a 5.4-kt iron-scintillator tracking calorimeter with a toroidal magnetic field, located underground in the Soudan mine at a distance of 735 km. The near detector is a small version of the far detector with a total mass of 0.98 kt. The neutrino beam is a horn-focused wide-band beam (see Section 17.7.1 of [1]). Its energy spectrum can be varied by moving the target position relative to the first horn and changing the horn current. MINOS obtained total of 3.36×10^{20} POT. In the detector, a total of 848 CC (charged current) events were produced by the NUMI beam compared to no oscillation expectation of 1065 ± 60 (syst) events. Figure 2.13b shows the MINOS data with oscillation hypothesis as well as alternative models (neutrino decay and decoherence).

The current best value of Δm_{32}^2 [16] is given as $2.32 \times 10^{-3}\ \text{eV}^2$. It means that the oscillation peak should be at

$$\begin{aligned} X_{32} &\equiv \frac{\Delta m_{32}^2 L}{4E} = 1.27 \frac{\Delta m_{32}^2\,(\text{eV})^2}{E\,(\text{GeV})} L\,(\text{km}) = \frac{\pi}{2} \\ \therefore \quad \frac{E\,(\text{GeV})}{L\,(\text{km})} &= 1.94 \times 10^{-3} \left(\frac{\Delta m_{32}^2}{2.41 \times 10^{-3}\ \text{eV}^2} \right) \end{aligned} \tag{2.59}$$

For $L = 250$ km (K2K) and 735 km (MINOS), $E = 0.48$ and 1.43 GeV, respectively. The dips in Figure 2.13 confirm these values.

OPERA: Unlike K2K and MINOS which measured reduction of the ν_μ flux, OPERA aims to confirm the atmospheric neutrino oscillation directly by detecting $\nu_\mu \to \nu_\tau$ conversion. The neutrino beam with an average energy of 17 GeV (CNGS, CERN neutrino to Gran Sasso) is produced at CERN by the PS (400-GeV PS). The far detector is placed at Gran Sasso in Italy at a distance of 730 km (Figure 2.10). An important ingredient of the detector is an emulsion target cum detector [183]. It is an age-old particle detection device used to record cosmic rays in early-day researches. As it records all the particles that pass through it, leaves unremovable tracks permanently, and no selection of events is possible, it has become obsolete as a particle detector, but here it is revived to take advantage of its capability to measure positions with near-micrometer accuracy.

The far detector (Figure 2.14), whose dimension is 10 m × 10 m × 20 m, is made of two identical super modules, each consisting of a 900-ton target section and a muon spectrometer.

The muon spectrometer behind the target section records the muon track and extrapolates back to find the interaction point in the target within $\lesssim 1\ \text{cm}^3$ volume. Then a microscope scanner finds the interaction point and secondary vertex to identify whether the muon is a decay product of the τ lepton.

Identification of the tau neutrino is done by reconstructing the following chain processes:

$$\nu_\tau + A \to \tau^- + X, \quad \tau \to \nu_\tau + \mu^- + \bar{\nu}_\mu \tag{2.60}$$

As the lifetime of the τ is $\sim 2.9 \times 10^{-13}$ s, decay flight distance is $\sim 100\ \mu\text{m}$. Detection of the primary interaction point and the secondary decay point is the signal for τ production (see lower illustration in Figure 2.14).

The target is a sandwich of emulsion/lead layers historically called the *ECC* (*emulsion cloud chamber*). The emulsion layer is a stacked wall of bricks with size

---

[16] Subfix "32" is attached for later reference when three-flavor oscillation is discussed.

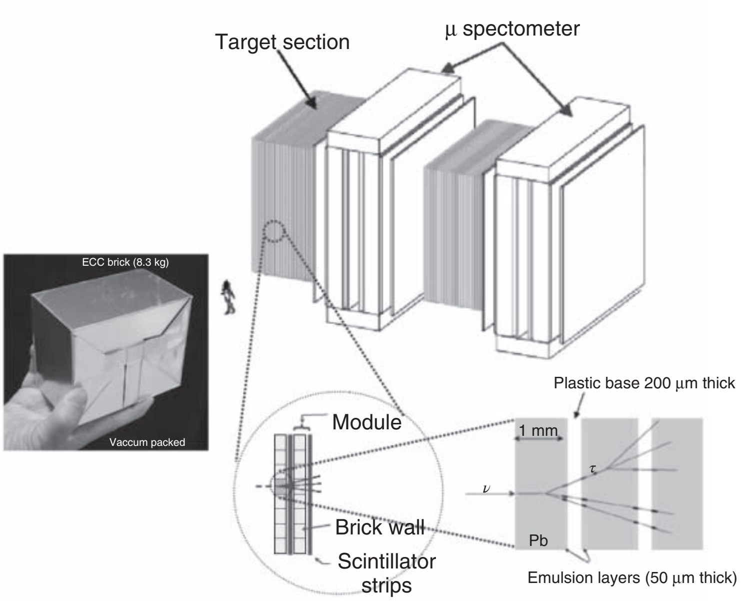


Figure 2.14 Illustration of the OPERA detector. It consists of two target and muon spectrometer sections. The muon spectrometer records muon tracks and extrapolates back to locate the interaction point in the target to within ~1 cm^3. Then, the emulsion scanning device finds and reconstructs the primary interaction and the secondary vertex points as illustrated. The left insert shows a picture of an emulsion brick. The total weight of the bricks is 1715 tons. (Reproduced with permission of [183].)

12.7 cm × 10.2 cm, and the thickness along the beam direction is 7.5 cm (about 10 radiation lengths) (see the inset picture in Figure 2.14). Each brick is a sandwich of fifty-six 1-mm-thick lead sheets and 57 emulsion sheets which are painted with 50 μm thickness on both sides of a 200-μm-thick plastic sheet. More than 200,000 bricks were assembled, and their total weight was 1715 ton.

Expecting a CNGS beam of 4.5×10^{19} POT per year, a five-year run would produce 31,000 CC events and ~150 tau events. Taking into account the detection efficiency, the group expects 10–15 signal events [184]. ν_τ events were observed. Their results are given in Refs [185, 186].

World data on the atmospheric neutrino oscillation: Contours of the allowed regions on the $\sin^2 2\theta - \Delta m^2$ plane obtained by the K2K and MINOS are shown in Figure 2.15 together with the SuperK result. The two accelerator experiments confirmed the result of the SuperK.

The present world average is given by [8]

$$\sin^2 2\theta_{atm} = \sin^2 2\theta_{32} > 0.95, \qquad |\Delta m^2_{atm}| = |\Delta m^2_{32}| = 2.32^{+0.12}_{-0.08} \times 10^{-3} \tag{2.61}$$

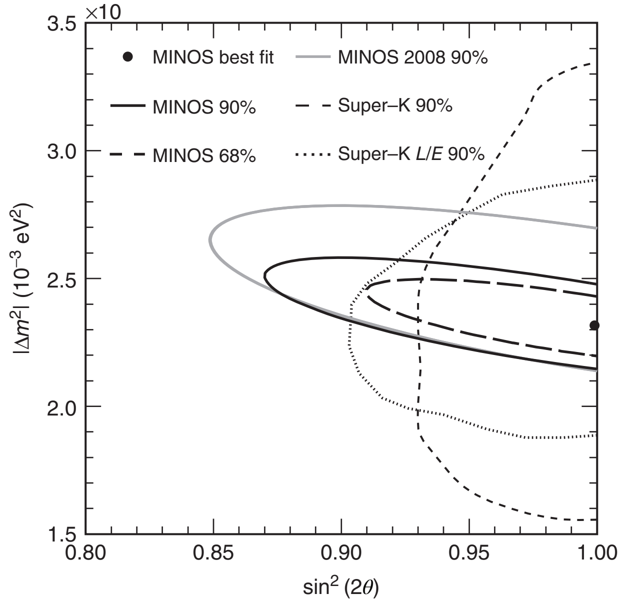


Figure 2.15 Allowed region for the $\nu_\mu \to \nu_\tau$ oscillation parameters from the MINOS and SK results. The 68% and 90% CL allowed regions are shown together with the SK 90% CL allowed regions. (Reproduced with permission of [8, 171].)

2.6
Underground Detectors

Cherenkov versus Scintillation Counter Observation of astrophysical neutrinos was pioneered by Davis, who started solar neutrino observation as early as 1968. A major breakthrough happened when the Kamiokande, a water Cherenkov detector, detected the neutrinos from Supernova 1987A and later from the Sun. Since then, many large underground detectors have been constructed. As they occupy an important position in particle physics, and their particle detection technology is somewhat different from that of accelerator experiments, we describe here essentials of a large-sized underground detector.[17] They are characterized by their large size, low-energy particle detection, and large environmental backgrounds.

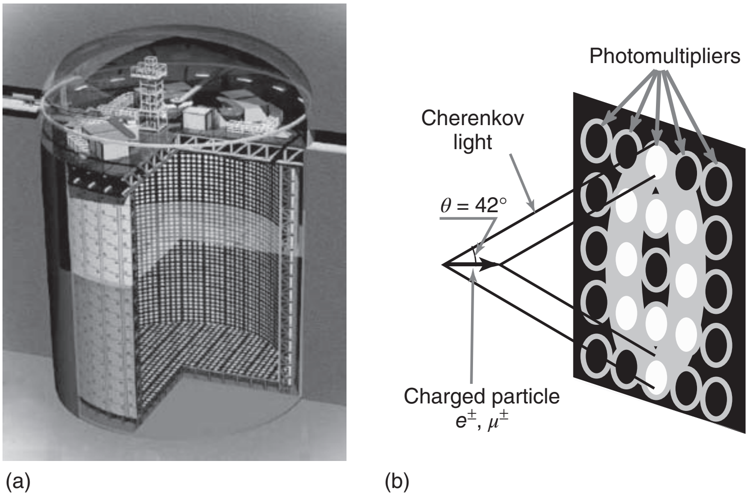
As typical neutrino detectors, we describe the detection mechanism of the SuperK and KamLAND detectors, one representing a water Cherenkov detector and the other representing a liquid scintillation detector. Description of the latter will be supplemented later in connection with the detection of reactor neutrinos.

Cherenkov light is characterized by its directionality, which can be used to reduce the background but also gives us confidence that we are actually observing the neutrinos from the desired source. Furthermore, Cherenkov light is insensitive to low-energy α particles which are abundantly produced in mineral radioactive materials. The Cherenkov detector is preferred for detecting electrons at higher energy ($\gtrsim$5 MeV) for which the photon yield is sufficiently large. On the other hand, if one wants to detect sub-MeV neutrinos, scintillation counters with high light yield are the unique choice. Because of the high statistics of the photon yield, energy resolution of $0.64{\sim}0.1/\sqrt{(E/\mathrm{MeV})}$ [187] can be obtained as compared to $0.5/\sqrt{(E/\mathrm{MeV})}$ for $E \lesssim 12$ MeV [188, 189] for the Cherenkov counter. When the coincidence technique cannot be used, as is the case for the solar neutrino observation, lack of directionality imposes a severe constraint on the reduction of the backgrounds. As a representative of giant underground detectors, we describe the SuperK detector.

---

[17] See also "Reactor experiment" in Section 2.7.3.

SuperKamiokande: The SuperK detector is placed at 1000 m (2700 m water equivalent (W.E.)) underground in the Kamioka Observatory of the Institute for Cosmic Ray Research, University of Tokyo, Gifu prefecture (see Figure 2.16a). It is a successor of the Kamiokande, which was originally designed as a detector for the proton decay predicted by GUTs and later converted to an astrophysical neutrino detector by reducing environmental backgrounds and lowering the threshold of detectable energy down to 5–6 MeV. The SuperK is a scaled-up Kamiokande (3→50 kt) with photomultipliers covering 40% of the surface of the water tank. The water tank has dimensions of $40\,\mathrm{m}\ \phi \times 42\,m$ height, and a total weight 50 kt. The inner detector (ID) is comprised of 11,146 Hamamatsu R3600 50-cm-diameter



(a)

(b)

Figure 2.16 (a) Cutaway view of the SuperK detector showing the inside lined with photomultiplier tubes (PMTs) comprising a photocathode coverage of about 40%. The thin outer region is shown with sparser density of outward-facing PMTs: two outer PMTs for every unit of 3 × 4 inner PMTs. The top of the detector, under the hemisphere, consists of electronics huts and open work area. (Reproduced with permission of [190].) (b) Detection principle. A charged particle scattered by a neutrino emits Cherenkov light which is detected by photomultipliers attached to the wall surface.

photomultiplier tubes(PMTs), viewing a cylindrical volume of pure water 16.9 m in radius and 36.2 m in height. The ID is surrounded by the outer detector (OD), which is a cylindrical shell of water 2.6–2.75 m thick including a dead space of 55 cm. The OD is optically isolated from the ID, and is instrumented with 1885 outward-facing Hamamatsu R1408 20 cm PMTs, providing both a 4π active veto and a thick, passive radioactivity shield. The information from the OD is used to identify both incoming and outgoing muons.

Event topology: The event topologies and their parent neutrino spectra are described in Figure 2.17. The fully contained events are further subdivided into sub-gigaelectronvolts and multi-gigaelectronvolts depending on the visible energy ($E_{\text{vis}} \gtrless 1.33$ GeV). The average energy of the incident neutrino is ∼1 GeV.

In the partially contained events, the scattered muons escape from the detector, and only the lower limit of the visible energy can be determined. The parent neutrino's energy is ∼10 GeV. The upward stopping and upward through-going muons are made by neutrinos coming from the other side of the Earth, interacting with nearby rocks, and entering the detector from below. The parent neutrino has much higher energy, ranging up to 10^4 GeV.

Particle identification: The principle of particle detection is depicted in Figure 2.16b. Scattered charged particles by the incoming neutrino emit Cherenkov lights at an angle of 42° relative to the flight direction of the parent particle. The vertex position is reconstructed from the hit time of PMTs and has a resolution of 30 cm for single-ring fully contained events. The particle identification is made by recognizing the ring pattern. The electron produces an electromagnetic shower and the ring shape is fuzzy, while the muon leaves a clear-edged ring (Figure 2.18).

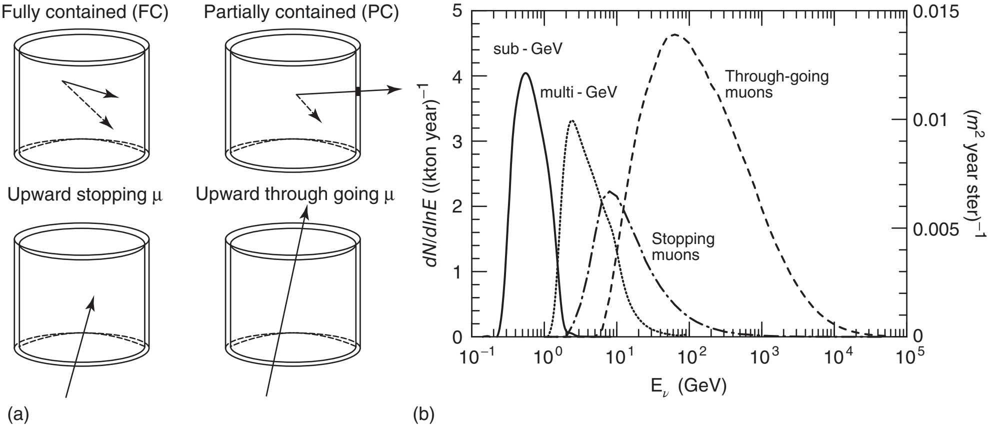


Figure 2.17 (a) Observed event topologies. (b) Parent neutrino energy distributions for the fully contained (FC, sub-GeV, and multi-GeV), upward stopping muon, and upward through-going muons samples [171, 191].

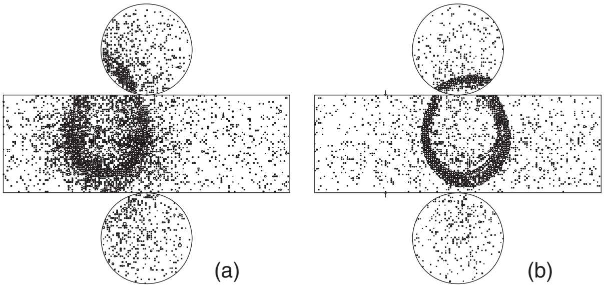


Figure 2.18 (a) Display of a single-ring *e*-like event. Each small circle represents a hit PMT, and the size of the circle represents the number of photons that hit it. In this event, the boundary of the Cherenkov light is smeared over many PMTs as the light comes from numerous positrons and electrons in the electromagnetic shower. (b) An example event display of a single-ring μ-like event. In this event, the boundary of the Cherenkov light is sharp because the muon travels relatively straight as it comes to a stop. Distant hit PMTs come from scattered light and Cherenkov light from delta-rays. (Reproduced with permission of [171].)

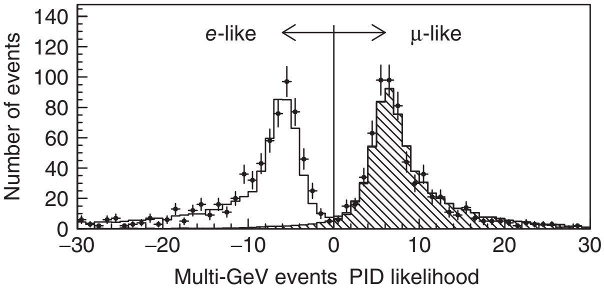


Figure 2.19 Distributions of particle identification likelihood for multi-GeV FC single-ring events, comparing SK data (points) and atmospheric neutrino Monte Carlo (histograms). The Monte Carlo includes neutrino oscillation with $\sin^2 2\theta = 1.00$, $\Delta m^2 = 2.1 \times 10^{-3}$ eV^2. The hatched histograms show the ν_μ charged-current interaction. (Reproduced with permission of [171].)

Validity of the particle identifying algorithm was confirmed by a beam test experiment at KEK [192]. Figure 2.19 shows a likelihood distribution for multi-GeV (visible energy >1.33 GeV) events. The misidentification probability for single-ring μ and electrons were estimated to be 0.7% and 0.8%, respectively.

Cosmic ray backgrounds: The major backgrounds are twofold: one produced by cosmic rays, and the other by natural radioisotopes in the nearby rocks and the detector materials themselves. The former can be reduced by going deep underground. Figure 2.20 shows the muon background intensity as a function of depth [190]. The SuperK detector is located at a depth of 1000 m, which amounts to 2700 m W.E. The background rate is reduced by 1/100,000 compared to the surface. Another factor 10 can be obtained by going to the depth 4000 m W.E. (Gran Sasso, Homestake). Muon tracks can be easily triggered out, but they leave excited nuclei

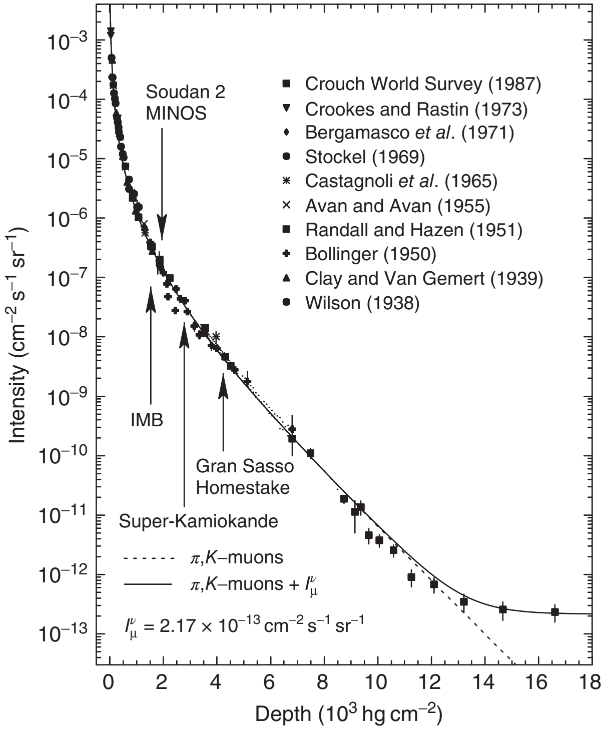


Figure 2.20 Muon background intensity as a function of the depth in units of water equivalent (W.E.). SuperKamiokande detector is located at 1000 m depth which is 2700 m W.E. The background rate is reduced by 1/100,000 compared to the surface. (Reproduced with permission of [190].)

by spallation (fission of the target oxygen or nucleosynthesis by emitted pions and slow neutrons) behind. Some examples are $^{8}B(\tau = 1.11$ s, $Q = 18.0$ MeV), $^{8}Li(\tau = 1.21$ s, $Q = 16.0$ MeV), $^{12}B(\tau = 29.1$ ms, $Q = 13.4$ MeV), $^{12}N(\tau = 15.9$ ms, $Q = 17.3$ eV) [193, 194]. The spallation events retain some correlation in time and space with their parent muons, and a cut (spallation cut) can be used to eliminate most of them. Some of the unstable nuclei are long-lived and are a source of steady background which becomes serious for triggerless signals (i.e., solar neutrinos) especially at low energies ($E \lesssim 6.5$ MeV).

Figure 2.21a shows a typical vertex distribution of low-energy events relevant for detection of the solar neutrino before the selection. Many of them of external origin (i.e., nearby rocks) are eliminated by the fiducial volume cut, as indicated by the dotted line, and by asking the directionality of the events (γ cut in SuperK), as most of them are strongly pointed inward. But some of external events still remain, which constitute the major backgrounds in this energy region. Figure 2.21b shows the number of events as a function of energy at each background reduction step. The final data sample is obtained after ambient background reduction by applying the γ cut and ^{16}N cut, which reduces the background generated by the capture of

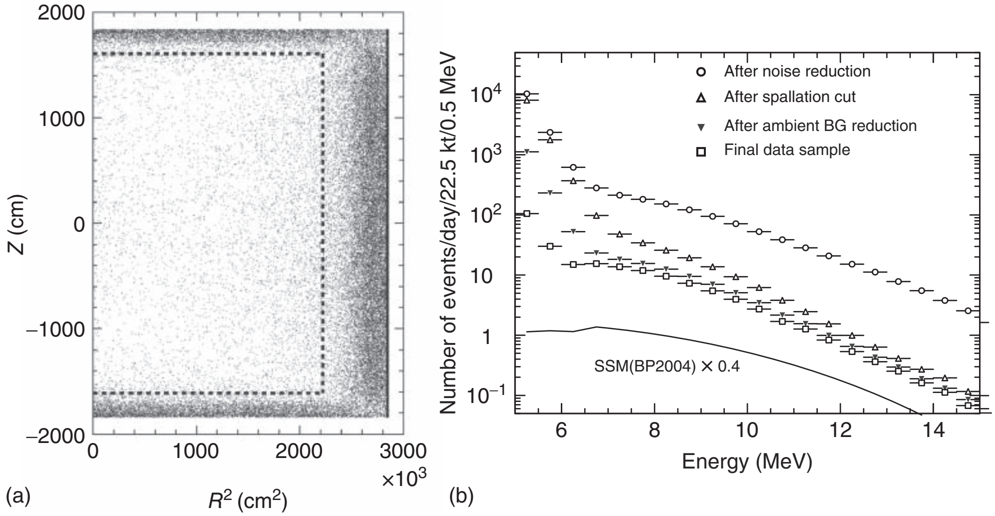


Figure 2.21 (a) Typical vertex distribution of the low-energy events before the intelligent trigger selection. The analysis energy threshold for this plot is 5.0 MeV. The dashed line shows the fiducial volume edge. (b) The event rate after each reduction step. The line shows the predicted solar neutrino spectrum. The final data sample is obtained after γ cut. The number of events after all the reduction steps is 286,557. (Reproduced with permission of [195].)

stopping μ^- on ^{16}O ($^{16}\text{O} + \mu^- \rightarrow {}^{16}\text{N} + \nu_\mu$). The final data sample is still an order of magnitude larger than the ^{8}B solar neutrino signal. The signal is extracted using the directionality of the $\nu_e e^-$ scattering (see Figure 2.26).

Internal backgrounds: After rejecting the external backgrounds from cosmic rays and surrounding rocks, internal backgrounds from radioactivity in the target material, air, the vessel, and the photomultipliers remain. They become more serious as the energy goes down.

The main agents are the natural isotopes including ^{238}U, ^{232}Th, ^{40}K, ^{14}C, ^{85}Kr, and ^{222}Rn. The concentration of ^{238}U in the Earth's crust or in normal water is typically ppm [parts per million = (10^{-6}g g^{-1})] or 12.4 Bq(Becquerel) l^{-1} (Bq = 1 disintegration/s).

In the spontaneous fission of uranium, energy is promptly released in the form of γ-rays with an energy of ~7 MeV, and the mean number of γ-rays is about 7.4. But the prime origin of the background is β-rays of 3.26 MeV from ^{214}Bi (used to be called *Ra C*). They are emitted in the $^{226}\text{Ra} \rightarrow {}^{222}\text{Rn}$ and $^{222}\text{Rn} \rightarrow \cdots \rightarrow {}^{214}\text{Bi} \rightarrow {}^{214}\text{Po}$ decay chain (see Figure 2.22) and are serious background sources. Purification steps of materials to reduce them include ion exchange, inverse osmosis, circulation of nitrogen, and so on.

The SuperK had to purify the water to reduce ^{222}Rn backgrounds typically to a level of 1.4 Bq kt^{-1} for successful observation of ^{8}B solar neutrinos. Subsequent detectors achieved further reduction, as listed in Table 2.2.

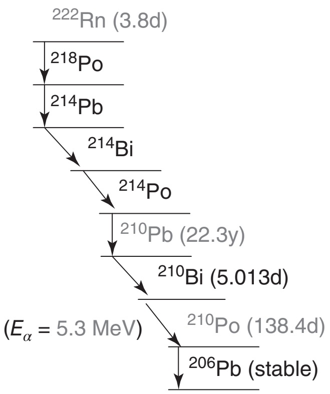


Figure 2.22 Source of major backgrounds contained in the detector materials. Starting with beta decay of ^{214}Bi, it includes ^{210}Pb which accompanies ^{210}Bi and ^{210}Po. The α-rays produced in the ^{210}Bi decay are one of the major background sources [$^{13}C(\alpha, n)^{16}O$] for geoneutrinos (see Figure 2.37).

Table 2.2 Inherent internal background in the detector. The units are Becquerel per kilotons (g g^{-1}).

	^{238}U	^{232}Th	^{222}Rn
Kamiokande [188, 196]	9	—	370
SuperK [197]	$<0.12\,(1 \times 10^{-14})$	$0.08\,(< 2 \times 10^{-14})$	1.4
SNO [198, 199]	$0.07\,(5 \times 10^{-15})$	$0.007\,(1.7 \times 10^{-15})$	0.04
Borexino [200]	$0.004\,(4 \times 10^{-16})$	$0.0018\,(4.7 \times 10^{-16})$	0.004
KamLAND [201]	$(3.5 \pm 0.5 \times 10^{-18})$	$(5.2 \pm 0.8 \times 10^{-17})$	3×10^{-6}

Below a few Megaelectronvolts, scintillation counters are used instead of Cherenkov counters. For detection of the ^{7}Be solar neutrino, which has an energy of 0.861 MeV and an event rate of 3×10^{-3} (kt^{-1} s^{-1}), the radioactivity in the scintillator must be at most 1×10^{-3} Bq kt^{-1}, a requirement 1000 times more severe than for the water Cherenkov counter.

2.7
Solar Neutrino

2.7.1
The Solar Puzzle

The Standard Solar Model (SSM) The source of solar energy is thermal fusion in the hot, pressurized plasma core of the Sun in which four protons are synthesized to form a helium atom and two positrons. The neutrino is a by-product of the fusion

$$2e^- + 4p \to {}^4\text{He} + 2\nu_e + \gamma\,(26.73\,\text{MeV}) \tag{2.62}$$

Its true identity is the electron neutrino ν_e, which is to be distinguished from the anti-neutrino $\bar{\nu}_e$ that is produced by reactors copiously. The thermal fusion is a result of multiple processes referred to as the *pp-chain* (Figure 2.23a).

If the core temperature exceeds 2×10^7 K, which is the case for stars heavier than the Sun, the carbon-nitrogen-oxygen (CNO) cycle in which C, N, and O catalyze the reaction becomes dominant (Figure 2.23b). The two processes have different energy spectra (Figure 2.24), which can be used to identify the origins of the observed neutrino.

Homestake Experiment The pioneers of the solar neutrino observation have been Davis and his group [203]. It dates back to 1968 [204]. The method was to use the reaction

$$\nu_e + ^{37}\mathrm{CL} \rightarrow ^{37}\mathrm{Ar} + e^- \tag{2.63}$$

which has a threshold energy $E_\nu > 0.81$ MeV. The target cum detector was a 615-ton liquid tetrachloroethylene (perchloroethylene) C_2Cl_4 (material used as detergent) tank, which was placed underground in the Homestake mine in South Dakota, USA, at a depth of 1620 m to reduce the cosmic ray background. Furthermore, the whole tank [picture shown in Figure 2.25] was immersed in a water pool to reduce the background due to slow neutrons.

The produced argon captures the orbiting electron and is transformed back to chlorine with half-life of 35 days. Therefore, if the detector works longer than that, the produced and captured numbers are in equilibrium and a steady number of argon atoms remain in the detector. The argon gas floating as bubbles in the liquid can be extracted by sending helium gas and adsorbing by active carbon. The argon emits a characteristic X-ray when it goes back to chlorine, and one can count its number by measuring it using proportional tubes. As the method depends on counting the number of argon atoms in the equilibrium state, one cannot know when the argon was captured nor from which direction it came. One simply assumes that all the detected neutrinos are of solar origin.

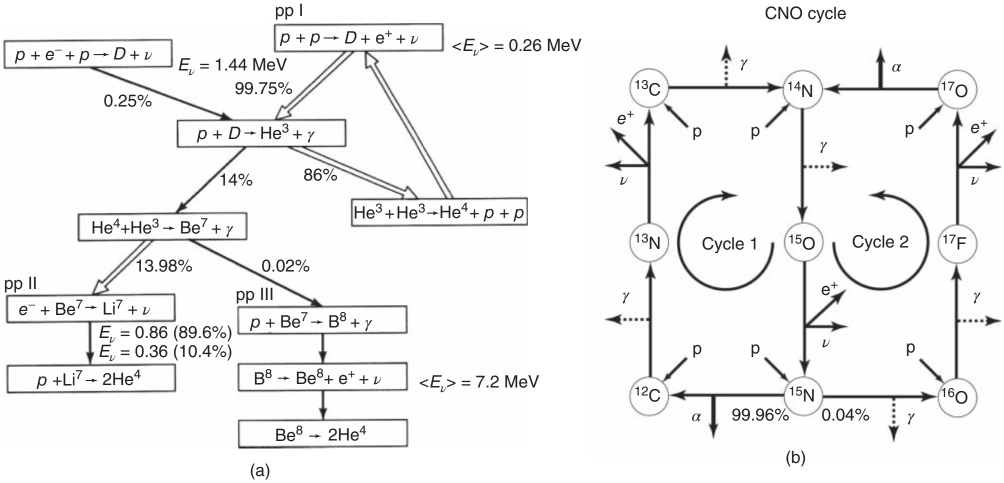
The Homestake group measured the average counting rate to be (Figure 2.25b)

$$2.33 \pm 0.25\ \mathrm{SNU}\ [205] \tag{2.64}$$

where SNU is the neutrino capture rate for 10^{36} Cl atoms. On the other hand, the SSM gives [202]

$$9.3^{+1.2}_{-1.4}\ \mathrm{SNU} \tag{2.65}$$

for the chlorine target. Therefore, the observed value is only one-third of what was predicted. Note that the theoretical uncertainty σ adopts 2σ of theoretical errors and is considered as conservative. When it was first presented in the 1970s, nobody could give a satisfactory explanation and remained as ‘‘the solar neutrino puzzle’’ for a long time. If one ascribes the cause to the solar model, it is a problem of


pp I
p + p → D + e⁺ + ν
<E_ν> = 0.26 MeV
p + e⁻ + p → D + ν
E_ν = 1.44 MeV
99.75%
0.25%
p + D → He³ + γ
14%
86%
He³+He³→He⁴+ p + p
He⁴+He³ → Be⁷ + γ
13.98%
0.02%
pp II
e⁻ + Be⁷→ Li⁷ + ν
E_ν = 0.86 (89.6%)
E_ν = 0.36 (10.4%)
p +Li⁷→ 2He⁴
pp III
p + Be⁷ → B⁸ + γ
B⁸ → Be⁸+ e⁺ + ν
<E_ν> = 7.2 MeV
Be⁸ → 2He⁴
CNO cycle
γ
α
¹³C
¹⁴N
¹⁷O
e⁺
p
p
p
e⁺
γ
ν
ν
¹³N
Cycle 1
¹⁵O
Cycle 2
¹⁷F
γ
e⁺
γ
p
p
ν
p
¹²C
¹⁵N
¹⁶O
α
99.96%
0.04%
γ

(a) (b)

Figure 2.23 Solar fusion consists of the pp chain (a) and the CNO cycle (b). The pp chain is the main fusion process in the solar core at a temperature $\sim 1.8 \times 10^7$ K. The CNO cycle becomes active above $\gtrsim 1.6 \times 10^7$ K. It produces only 1.6% of the total energy in the Sun but is dominant in the large mass stars.

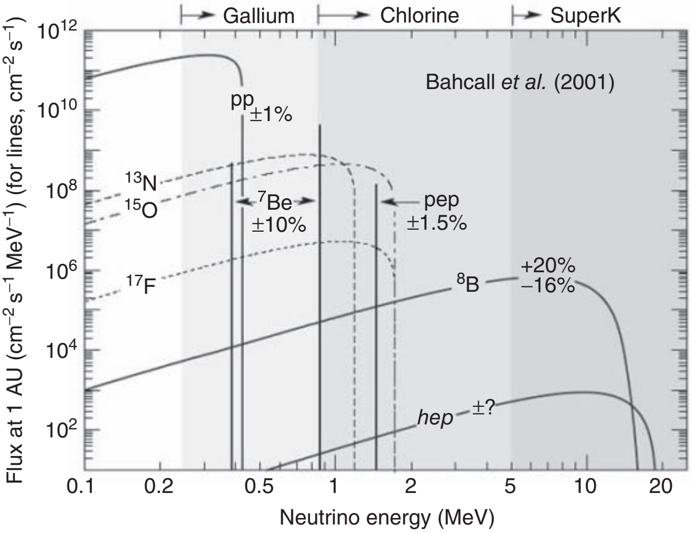
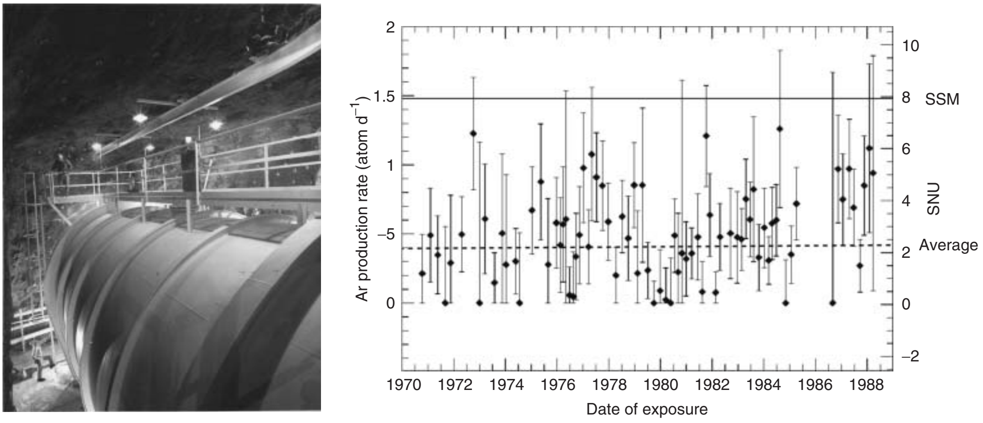


Figure 2.24 Solar energy spectrum due to pp-chain and CNO cycle. The threshold energies for different detectors are indicated on top of the figure. (Reproduced with permission of [202].)



(a) (b)

Figure 2.25 (a) Davis's detector at the Homestake mine. A 615-ton liquid tetrachloroethylene (C_2Cl_4) tank was placed underground in the Homestake mine at a depth of 1620 m. To reduce the cosmic ray background, the whole apparatus is immersed in a pool of water. Photo courtesy of R. Davis. (b) Neutrino reaction rate per day plotted as a function of time (year). The unit is SNU/day (SNU (solar neutrino unit) = neutrino capture rate per 10^{36} Cl atoms). (Reproduced with permission of [204, 203].)

astrophysics, and if one ascribes it to the property of the neutrino itself, it is a problem of particle physics. One astrophysical explanation was that the solar core temperature was low some hundreds of millions years ago because the energy transport from the core to the surface takes that much time. Another was that there are more heavy elements than imagined, thereby reducing the reaction rate. There were many more. However, the SSM is based on the stellar evolution theory, which is supported by a great number of star observations. One small change of the standard inputs invites some conflicts with observations. There were blames on the experiment itself, but the result has survived the long trial of time.

SuperKamiokande In the 1980s, there were experiments to measure the proton lifetime inspired by GUTs. Those devices were also ideal neutrino detectors with due improvements. The Kamiokande detector was one of them, and made a historical observation of the supernova 1987A together with the IMB detector. Although it also detected the solar neutrino, an accurate measurement was done by its successor SuperK.

Detection of particle reactions was made by catching Cherenkov light that is emitted by the scattered charged particles in water. From the number and the pattern of the Cherenkov ring, the detector can identify electrons, muons, and π^0's. From the amount of light, the energy can be measured. The mentioned improvement to convert the proton decay detector to the solar neutrino detector was to reduce the background of natural environmental origin and to lower the detection energy threshold down to $7-8$ MeV. The reduction of the environmental background was done by adding an OD to shield the ID from the background.

The reduction of the energy threshold is not so critical for proton decay because the proton liberates a total energy of 940 MeV. But it was essential for the solar neutrino observation where the average solar neutrino energy is $\lesssim$1 MeV with maximum energy barely going to $10-15$ MeV (see Figure 2.24). The low-energy neutrino can be detected when it scatters electrons elastically in the water medium

$$\nu_e + e^- \to \nu_e + e^- \tag{2.66}$$

and when the scattered electron emits Cherenkov light. As the direction of the scattered electron is the same as that of the incoming neutrino to a good precision,[18] one can measure its arrival time, flight direction, and energy spectrum. In that sense, it is a more advanced detector, which may qualify as a neutrino telescope.

The solar neutrino reactions are rare events, and no coincidence method was available to select the desired events. It was essential to reduce the background emitted by isotopes of uranium and other decay chains that are abundant in the surrounding rocks. The background increases as the energy goes down (see Figure 2.21).

Figure 2.26a shows the accumulated number of $\nu_e e$ reaction events as a function of the neutrino flight direction [195, 206]. The direction of the Sun corresponds to

[18] The angle θ and energy E of the scattered electron are related by $\theta^2 < 2m_e/E$ for $E \gg m_e$ (see Section 2.2 of [2]).

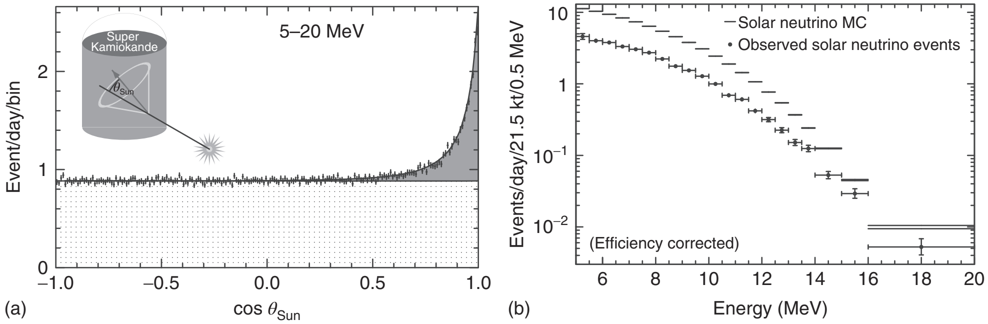


Figure 2.26 (a) Solar neutrino data accumulated over 1496 days. Left: as a function of $\cos\theta_{\text{Sun}}$ where $\cos\theta_{\text{Sun}} = 1$ is the direction of the sun. (b) Energy spectrum. Its shape is the same as the SSM [207] but with much smaller absolute values. (Reproduced with permission of [208, 195, 206].)

$\cos\theta_{\text{Sun}} = 1.0$. One can see clear signals in that direction. Figure 2.26b shows their energy spectrum.

The observed flux by the SuperK is measured to be

$$N_\nu = 2.44 \pm 0.05\,(\text{stat})^{+0.09}_{-0.07}\,(\text{sys}) \times 10^6\ \text{cm}^{-2}\ \text{s}^{-1}\ (E_e \geq 6.5\ \text{MeV}) \tag{2.67a}$$

$$\frac{N(\text{observed})}{N(\text{theory})} = 0.474^{+0.010+0.017}_{-0.009-0.014} \tag{2.67b}$$

The observed value is about one-half of the predicted value [209].[19] Although the ratio is different from that obtained by Davis, which can be ascribed to the different energy sensitivity of the detector, the fact that the observed value is much smaller than predicted was confirmed.

Ga Experiment The two experiments discussed so far are only sensitive to the high-energy part of the solar neutrino flux. Note that the arrows at the upper part of Figure 2.24 denote the sensitive region. This is also the region where estimate of the SSM is somewhat ambiguous. Therefore, it becomes important to measure the solar flux in the low-energy region where the pp-I chain, the dominant contribution to the flux, is active. If one uses gallium instead of chlorine and applies the same method as the Homestake experiment, one can measure the neutrino from the pp-I chain because the reaction ${}^{71}\text{Ga}(\nu_e, e^-){}^{71}\text{Ge}$ has a much lower threshold ($E \geq 233$ keV). Two groups, GALLEX/GNO (Gallium Experiment/Gallium Neutrino Observatory) in Italy and SAGE (Soviet-American Gallium Experiment) in Russia, carried out

---

[19] Later, events at lower energies ($5.0 \leq E_e \leq 6.5$ MeV) were added, and the final value is [195].

$$N_\nu = 2.35 \pm 0.02\,(\text{stat}) \pm 0.08\,(\text{sys}) \times 10^6\ \text{cm}^{-2}\text{s}^{-1}\ (E \geq 5.0\ \text{MeV}) \tag{2.68}$$

these experiments. The measured values are summarized below:

$$\begin{array}{lllll} \text{GALLEX}: & 77.5 & \pm 6.2\,(\text{stat}) & {}^{+4.3}_{-4.7}\,(\text{syst}) & \text{SNU [210]} \\ \text{GNO}: & 62.9 & {}^{+5.5}_{-5.3}\,(\text{stat}) & \pm 2.5\,(\text{syst}) & \text{SNU [211]} \\ \text{SAGE}: & 65.4 & {}^{+3.1}_{-3.0}\,(\text{stat}) & {}^{+2.6}_{-2.8}\,(\text{syst}) & \text{SNU [212]} \\ \text{Combined}: & 66.1 & \pm 3.1 & & \text{SNU [212]} \end{array} \tag{2.69}$$

which is to be compared with the SSM prediction 129^{+8}_{-6} SNU [202].

One sees that the solar puzzle is firmly established.[20]

SNO Experiment The definitive proof of the solar puzzle due to the neutrino oscillation was presented by the Sudbury Neutrino Observatory (SNO) group [198, 169, 199]. The detector was 1000 tons of heavy water with or without salt which enhances the neutron capture rate. Three kinds of event topologies were used.

$$\begin{aligned} \text{CC}: &\quad \nu_e + D \to p + p + e^- - 1.442\,\text{MeV} \\ \text{NC}: &\quad \nu_x + D \to p + n + \nu_x - 2.224\,\text{MeV} \\ \text{ES}: &\quad \nu_x + e^- \to \nu_x + e^- \end{aligned} \tag{2.70}$$

where ν_x refers to any active flavor of neutrinos and CC, NC, and ES stand for charged current reaction, neutral current reaction, and elastic scattering.[21] The essence of the experiment is to detect both CC and NC reactions at the same time, using the same target and the same detector. As Eq. (2.70) show, CC events catch ν_e only, but NC and ES are sensitive to all types of flavors. Note that NC couples to all flavors with equal strength, but ES is sensitive primarily to ν_e because of the $W^\pm$ exchange process. The Standard Model constrains them to satisfy

$$\begin{aligned} \phi_{\text{CC}} &= \phi_e \\ \phi_{\text{NC}} &= \phi_e + \phi_{\mu\tau} \\ \phi_{\text{ES}} &= \phi_e + 0.1553\phi_{\mu\tau} \end{aligned} \tag{2.71}$$

where $\phi_{\text{CC}}, \phi_{\text{NC}}, \phi_{\text{ES}}$ are the neutrino fluxes determined by the processes in Eq. (2.70), and ϕ_e and $\phi_{\mu\tau}$ are the fluxes of ν_e and sum of $\nu_\mu + \nu_\tau$.

If the solution to the solar puzzle is the neutrino oscillation in which a fraction of ν_e flux is converted to ν_μ, ν_τ, the total flux determined by the NC process should be larger than that determined by CC. The ES process can be used to test the consistency of the hypothesis given by three constraints of the Standard Model. The measurement by the SNO produced [198, 169, 199]

$$\begin{aligned} \phi_{\text{CC}} &= 1.68^{+0.06}_{-0.06}\,(\text{stat})^{+0.08}_{-0.09}\,(\text{syst}) \\ \phi_{\text{ES}} &= 2.35^{+0.22}_{-0.22}\,(\text{stat})^{+0.15}_{-0.15}\,(\text{syst}) \\ \phi_{\text{NC}} &= 4.94^{+0.21}_{-0.21}\,(\text{stat})^{+0.38}_{-0.34}\,(\text{syst}) \end{aligned} \tag{2.72}$$

---

[20] The results given here are recent ones, long after the neutrino oscillation was established. Therefore, the data were used to confirm the validity of both the SSM and the neutrino oscillation.

[21] In the Standard Model, there are only three flavors. Experimentally, however, any neutrino that can participate in the reaction has to be considered and is denoted as an active neutrino.

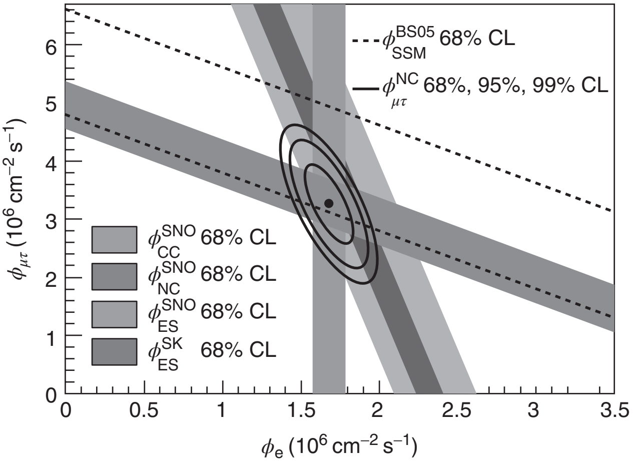


Figure 2.27 SNO data consistency check. The solar neutrino fluxes are measured by three different reactions CC : $\nu_e + D \to e^- + p + p$, NC : $\nu_e + D \to \nu_e + p + n$, and ES : $\nu_e + e^- \to \nu_e + e^-$. Three bands are the SM constraints that express $\phi_{\mu\tau}$ (sum of ν_μ and ν_τ fluxes) as functions of ϕ_e (ν_e flux). The band between the two dotted lines is the prediction of the SSM. The narrow black band of ϕ_{ES} indicates the νe scattering data obtained by SuperK. (Reproduced with permission of [7].) (Please find a color version of this figure on the color plates.)

Apparently, $\phi_e = \phi_{\text{CC}}$ is smaller than other fluxes, which include contributions of the other neutrinos. This is a firm evidence of neutrino oscillation.

For the consistency check, one substitutes Eq. (2.72) in Eq. (2.71), which define three lines in the $\phi_e - \phi_{\mu\tau}$ plane, and see whether they meet at one point. Indeed, they do meet and confirm the Standard Model and the hypothesis of neutrino oscillation. Furthermore, ϕ_{NC} agrees with the total solar neutrino flux predicted by the SSM, which is indicated by the band between the two dotted lines in Figure 2.27. The long-standing solar puzzle, which dates back to as early as 1970, was solved in 2002 [169].

2.7.2
Matter Oscillation

Schrödinger Equation in Matter We pointed out earlier that Earth is transparent to the neutrino in explaining the zenith angle distribution of the atmospheric neutrino. However, this is not quite true. Excluding the ubiquitous but negligible gravity, the neutrino interacts only weakly. Earth is just too thin for the neutrino to show up. That is the conventional thinking. An eye-opening proposal was made by Mikheyev and Smirnov [213] that the neutrino behaves differently in matter. They showed the possibility that, even if the mixing is as small as 10^{-3}, the matter effect of the Sun will cause a resonant oscillation, possibly explaining the solar puzzle. Note that this was before SNO's experiment, but even after settlement of the solar puzzle it played a crucial role in determining the mixing parameters correctly. The matter effect appears, in simple words, as the refraction just like that of light in matter. It is the effect of the forward scattering of the photon in matter added coherently with incoming waves, which changes the wavelength (i.e., momentum) of the light. Similarly, the forward scattered ν_e interferes coherently with unscattered ν_e and changes its wavelength (or effectively its mass), which in turn changes the oscillation wavelength.

Let us first consider the Schrödinger equation which the mass eigenstate of the neutrino satisfies in vacuum.

$$i\frac{d|\nu_j(t)>}{dt} = \left(p + \frac{m_j^2}{2p}\right)|\nu_j(t)> \tag{2.73}$$

Let us express the mass matrix corresponding to the second term on the right-hand side of Eq. (2.73) in the representation of $(|\nu_e>, |\nu_\mu>)$ bases which are given by Eq. (2.45). Denoting the wave state vector of the two-flavor neutrinos as Ψ, its equation of motion can be derived from Eqs. (2.45) and (2.46), and is written as

$$i\frac{d\Psi}{dt} = \mathscr{M}\Psi, \qquad \Psi = \begin{bmatrix}\alpha(t)\\ \beta(t)\end{bmatrix} \equiv \alpha(t)|\nu_e(0)> + \beta(t)|\nu_\mu(0)> \tag{2.74}$$

where the mass matrix $\mathscr{M}$ is defined by

$$\mathscr{M} = M + \begin{bmatrix}M_{ee} & M_{e\mu}\\ M_{\mu e} & M_{\mu\mu}\end{bmatrix} = M + \frac{\Delta m^2}{4p}\begin{bmatrix}-\cos 2\theta & \sin 2\theta\\ \sin 2\theta & +\cos 2\theta\end{bmatrix} \tag{2.75a}$$

$$M = p + \frac{1}{2p}\frac{m_1^2 + m_2^2}{2}, \qquad \Delta m^2 = m_2^2 - m_1^2 > 0 \tag{2.75b}$$

$$\tan 2\theta = \frac{2M_{e\mu}}{M_{\mu\mu} - M_{ee}} \tag{2.75c}$$

The first term M does not contribute to the oscillation. These are the equations that hold in vacuum. In matter, the above equations are modified as a result of interactions of ν_e and ν_μ with matter. The effect of the NC interaction on them is equal, thereby changing the value of the mass M but not the mass difference (Δm^2), and hence does not change the oscillation process. On the other hand, the CC interaction induces $\nu_e e$ scattering but leaves the ν_μ unaffected. The forward scattering with matter can act coherently and can be taken into account in terms of refractive index n. Then the plane-wave state of the neutrino $\sim e^{ipx-iEt}$ changes its momentum $p \to np$. In order to see the difference from the original wave, we rewrite $ct \to x$ and express the wave as

$$|\nu(x)>= |\nu>\, e^{inpx-iEx} = e^{ipx}e^{-i[-(n-1)p+E]x} \tag{2.76}$$

Using the known formulas in optics [see Eq.(16.32) of [1]], the refractive index can be expressed as

$$p(n-1) = \frac{2\pi n_e}{p}f(0) = -\sqrt{2}G_F n_e \tag{2.77}$$

where n_e is the electron number density, $f(0)$ is the forward $\nu_e - e$ scattering amplitude, and G_F is the Fermi coupling constant. For $\overline{\nu}_e e$ scattering, sign of $f(0)$

has to be changed. Because of Eq. (2.77), M_{ee} in the mass matrix is modified to

$$M_{ee} \to M_{ee} + \sqrt{2}G_F n_e \tag{2.78}$$

As $E \approx$p, we use E instead of p in the following. If one regards n_e as constant, the mass eigenstates can be solved just like those in vacuum by diagonalizing the mass matrix. Setting the energy eigenvalues and mixing angle in matter as $p + \tilde{m}_\nu^2/2E$ and θ_m, respectively, they are given by

$$p + \frac{\tilde{m}_\nu^2}{2E} = M + \frac{G_F n_e}{\sqrt{2}} \pm \frac{\Delta m^2}{4E}\left[\left(\cos 2\theta - \frac{2\sqrt{2}G_F n_e E}{\Delta m^2}\right)^2 + \sin^2 2\theta\right]^{1/2}$$

$$\therefore \quad \Delta\tilde{m}_\nu^2 = \Delta m^2 \left[\left(\cos 2\theta - L_\nu/L_r\right)^2 + \sin^2 2\theta\right]^{1/2} \tag{2.79a}$$

$$\sin 2\theta_m = \frac{\sin 2\theta}{\sqrt{\sin^2 2\theta + (\cos^2 2\theta - L_\nu/L_r)^2}} \tag{2.79b}$$

$$L_\nu = \frac{4\pi E}{\Delta m^2}, \qquad L_r = \frac{\pm 2\pi}{\left(\sqrt{2}G_F n_e\right)} \tag{2.79c}$$

where L_ν is the oscillation wavelength in vacuum and the ± sign for L_r is for the neutrino/anti-neutrino. Substituting $G_F = 1.436 \times 10^{-49}$ erg cm^3, and restoring $\hbar$ and c

$$L_r = 2\pi\frac{\hbar c}{\sqrt{2}G_F n_e} \approx 3.25 \times 10^7\,[\rho\,(\mathrm{gr/cm^3})]^{-1}\,\mathrm{m} \tag{2.80}$$

where ρ is the density of the matter. One can approximate the electron number density in the Sun by the equation [214]

$$\begin{aligned} &n_e = n_{ec} e^{-\frac{r}{R_0}} \\ &n_{ec} \approx 98.6 N_A \mathrm{cm}^{-3}, \quad N_A, \text{Avogadro number}, \\ &R_0 \approx R_\odot/10 \approx 7 \times 10^7\,\mathrm{m} \end{aligned} \tag{2.81}$$

where $R_\odot$ denotes the solar radius. The density ρ is $\sim$100 g/cm^{-3} at the center, but the typical value (say at $r \simeq 0.4R_\odot$) is $\sim$2 g/cm^{-3}. The neutrino mass and the mixing angle in the Sun are plotted in Figure 2.28 as a function of the electron number density n_e.

If $m(\nu_e)$ is lighter than $m(\nu_\mu)$ in vacuum, the matter effect increases the mass of ν_e and it becomes larger when $n_e > n_{er}$, where n_{er} is the density at crossing. If there is no mixing ($\sin 2\theta = 0$), they are expressed by two dotted lines which cross at $n_e = n_{er}$. However, the mixing separates the two states and the eigenstates in matter are expressed by two real lines. If the density at the core is larger than that at the crossing ($n_{ec} > n_{er}$), ν_e produced at the core is almost $\tilde{\nu}_2$. As the neutrino departs from the core, the density decreases and approaches the crossing point. If the mixing is maximal (the resonant mixing), the $\tilde{\nu}_2$ line reaches the exit where $n_e = 0$ without crossing the $\tilde{\nu}_1$ line. At the exit, $\tilde{\nu}_2$ is in pure ν_μ state. That is, ν_e is

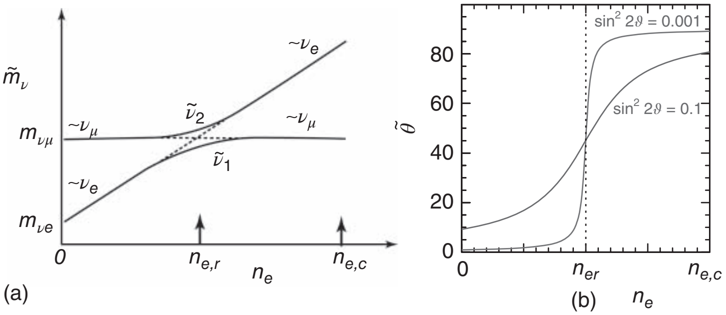


Figure 2.28 (a) Neutrino mass in the Sun plotted as a function of the solar density. The mass eigenstates are expressed by two solutions $\tilde{\nu}_1, \tilde{\nu}_2$. However, at the core where $n_e >> n_{er}$, $\tilde{\nu}_2 \sim \nu_e$, and at periphery where $n_e << n_{er}$, $\tilde{\nu}_2 \sim \nu_\mu$. Therefore, ν_e produced at the core is almost $\tilde{\nu}_2$, and comes out as ν_μ when it exits the Sun. (b) Mixing angle θ_m (denoted as $\tilde{\theta}$ here) plotted as a function of the number density n_e.

converted to ν_μ completely when it exits the Sun. This is referred to as the *MSW* (*Mikheyev-Smirnov-Wolfenstein*) effect [213, 215, 216].[22)]

The two lines are closest at $n_e = n_{er}$, which translates to $\cos 2\theta \sim L_\nu/L_r$ by Eq. (2.79a). This is where the mixing is maximal, that is, $\tan\theta_m = \infty$ (resonance). One sees that, even if the mixing angle θ in vacuum is very small, the mixing in matter can be maximal if the condition is favorable. The resonance occurs at

$$\theta_m = \frac{\pi}{4} \;\rightarrow\; \cos 2\theta = \frac{L_\nu}{L_r} = \frac{2\sqrt{2}G_F E n_e}{\Delta m^2}$$
$$\text{or}\quad n_e = n_{er} \equiv \frac{\Delta m^2 \cos 2\theta}{2\sqrt{2}G_F E} \simeq 6.56\times 10^6 \frac{\Delta m^2(\text{eV})^2}{E(\text{MeV})} N_A \cos 2\theta\,\text{cm}^3 \tag{2.82}$$

where n_{er} is called (for $\Delta m^2 \cos 2\theta > 0$) the *resonance density*, N_A being the Avogadro number. Equation (2.82) can be rewritten in terms of the mass squared difference as

$$\Delta m^2 \cos 2\theta\,(\text{eV})^2 = 1.52\times 10^{-7}\left(\frac{E_\nu}{\text{MeV}}\right)\left(\frac{\rho}{\text{g cm}^{-3}}\right) \tag{2.83}$$

Considering $E_\nu \lesssim 10$ MeV and $\rho < 100\,\text{g cm}^{-3}$, the oscillation can be significant in the mass range

$$10^{-8} < \Delta m^2\,(\text{eV})^2 \lesssim 10^{-4} \tag{2.84}$$

where the lower limit comes from the adiabatic condition [Eq. (2.90)]. One sees that the sensitive area where one can detect the finite oscillation amplitude is considerably expanded compared to the vacuum oscillation.

[22)] Wolfenstein constructed the matter oscillation formalism [217, 218] long before MS's idea was proposed.

The resonant mixing occurs when the density at the core is higher than that at the resonance point. Since the core density is given by Eq. (2.81), the condition ($n_{ec} \geq n_{er}$) for the resonance to occur in the Sun is expressed as

$$\Delta m^2 \cos 2\theta \lesssim 1.5 \times 10^{-4} \left(\frac{E}{10\,\text{MeV}} \right) \tag{2.85}$$

Adiabatic Approximation Arguments in the previous section are a bit deceptive. We assumed a constant density to solve the Schrödinger equation, yet we talked about the mass variation as a function of the density. The truth is, if the density change is sufficiently slow compared to the time variation of the oscillation, the adiabatic approximation can be applied. It means that, provided the above assumption is valid, the time variation of the solution can be obtained by introducing the time variation of the density to the solution obtained with the constant density assumption. To quantify the adiabatic condition, we require that the energy change δE due to the density variation be much smaller than the energy difference ΔE between the two eigenstates. In other words, the necessary distance L_ρ for δE to reach ΔE must be sufficiently long compared to the oscillation wavelength in matter $L_m/2\pi = 2E/\Delta\tilde{m}^2$. From Eq. (2.78), the energy change due to the density change is given by $\delta E = \sqrt{2} G_F \delta n_e$. Equating δE obtained by travelling the distance L_ρ with ΔE, we can calculate L_ρ.

$$\sqrt{2} G_F \frac{dn_e}{dx} L_\rho = \Delta E = \frac{\Delta\tilde{m}^2}{2E} \tag{2.86}$$

$$\therefore \quad L_\rho = \frac{\Delta\tilde{m}^2/2E}{\sqrt{2} G_F (dn_e/dx)} \tag{2.87}$$

Consequently, the adiabatic condition becomes

$$\gamma_m \equiv \frac{L_\rho}{L_m/2\pi} = \frac{(\Delta\tilde{m}^2)^2}{4\sqrt{2} G_F E^2 (dn_e/dx)} \gg 1 \tag{2.88}$$

At the resonance point

$$\Delta\tilde{m}^2 = \Delta m^2 \sin 2\theta, \qquad n_e = n_{er} = \frac{\Delta m^2 \cos 2\theta}{2\sqrt{2} G_F E} \tag{2.89}$$

and the adiabatic condition can be rewritten as

$$\gamma_m = \frac{\Delta m^2 \sin^2 2\theta}{2E \cos 2\theta (d \ln n_e/dx)} \gg 1 \tag{2.90}$$

Substituting Eq. (2.81) in Eq. (2.90), we finally obtain

$$\left(\frac{\Delta m^2}{1\,\text{eV}^2} \right) \sin 2\theta \tan 2\theta \gg 5 \times 10^{-8} \left(\frac{E}{10\,\text{MeV}} \right) \tag{2.91}$$

Survival Probability In the adiabatic approximation solution, the mass eigenstates $\tilde{\nu}_1$ and $\tilde{\nu}_2$ do not mix and the complete conversion $\nu_e \to \nu_\mu$ takes place. In general, however, the transition $\tilde{\nu}_2 \to \tilde{\nu}_1$ occurs at the resonance region. When the transition probability P_{jump} is given, the survival probability of ν_e can be obtained as follows:

Let us assume that ν_e is produced at the center of the Sun at $t = 0$. In order to expand $|\nu_e(t)>$ in the mass eigenstates, we replace $\theta \to \theta_m(t = 0) \equiv \tilde{\theta}(t = 0)$ in Eq. (2.45).

$$|\nu_e(t = 0)> = \cos\tilde{\theta}|\tilde{\nu}_1> + \sin\tilde{\theta}|\tilde{\nu}_2> \tag{2.92}$$

Setting $t = t_r$ when the neutrino arrives at the resonance region, the time evolution of the state for $t \le t_r - \varepsilon$ is given by

$$|\nu_e(t)> = \cos\tilde{\theta}\exp\left(-i\int_0^{t_r-\varepsilon} dt\,\tilde{E}_1\right)|\tilde{\nu}_1> + \sin\tilde{\theta}\exp\left(-i\int_0^{t_r-\varepsilon} dt\,\tilde{E}_2\right)|\tilde{\nu}_2> \tag{2.93}$$

where $\tilde{E}_1,\ \tilde{E}_2$ are the energy eigenvalues in matter. The transition between $|\tilde{\nu}_1> \leftrightarrow |\tilde{\nu}_2>$ occurs at $t_r - \varepsilon \le t \le t_r + \varepsilon$. We write down the states after the transition as

$$|\tilde{\nu}_1> \to \quad \alpha|\tilde{\nu}_1> + \beta|\tilde{\nu}_2> \tag{2.94a}$$

$$|\tilde{\nu}_2> \to -\beta^*|\tilde{\nu}_1> + \alpha^*|\tilde{\nu}_2> \tag{2.94b}$$

$$|\alpha|^2 + |\beta|^2 = 1 \tag{2.94c}$$

Then, $|\nu_e(t)>$ becomes

$$\begin{aligned}|\nu_e(t)> &= \cos\tilde{\theta}\exp\left(-i\int_0^{t_r-\varepsilon} dt\,\tilde{E}_1\right)(\alpha|\tilde{\nu}_1> + \beta|\tilde{\nu}_2>)\\ &\quad + \sin\tilde{\theta}\exp\left(-i\int_0^{t_r-\varepsilon} dt\,\tilde{E}_2\right)(-\beta^*|\tilde{\nu}_1> + \alpha^*|\tilde{\nu}_2>)\\ &\equiv A|\tilde{\nu}_1> + B|\tilde{\nu}_2>\end{aligned} \tag{2.95}$$

For $t \ge t + \varepsilon$, the evolution is similar to Eq. (2.93), but at $t \to \infty$, the neutrino exits the Sun and the matter eigenstates go back to those in vacuum.

$$|\tilde{\nu}_1>,\ |\tilde{\nu}_2> \xrightarrow{t\to\infty} |\nu_1>,\ |\nu_2> \tag{2.96}$$

Using Eq. (2.45) again to rewrite $\nu_e(t)$ in terms of vacuum states $\nu_e,\ \nu_\mu$, we have

$$\begin{aligned}|\nu_e(t = \infty)> &= A\exp\left(-i\int_{t_r+\varepsilon}^{\infty} dt\,\tilde{E}_1\right)(\cos\theta|\nu_e> - \sin\theta|\nu_\mu>)\\ &\quad + B\exp\left(-i\int_{t_r+\varepsilon}^{\infty} dt\,\tilde{E}_2\right)(\sin\theta|\nu_e> + \cos\theta|\nu_\mu>)\end{aligned} \tag{2.97}$$

Then the survival amplitude becomes

$$<\nu_e|\nu_e(t)> = A\cos\theta\exp\left(-i\int_{t_r+\varepsilon}^{\infty} dt\,\tilde{E}_1\right) + B\sin\theta\exp\left(-i\int_{t_r+\varepsilon}^{\infty} dt\,\tilde{E}_2\right) \tag{2.98}$$

which leads to the survival probability

$$P(\nu_e \to \nu_e) = |A|^2\cos^2\theta + |B|^2\sin^2\theta \tag{2.99}$$

The interference term vanishes after averaging over finite energy resolution, and spreads in production as well as in detection points. Substituting Eq. (2.95) in

Eq. (2.99), we finally obtain the survival probability for the ν_e as

$$P(\nu_e \to \nu_e) = (|\alpha|^2 \cos^2 \tilde{\theta} + |\beta|^2 \sin^2 \tilde{\theta}) \cos^2 \theta + (|\beta|^2 \cos^2 \tilde{\theta} + |\alpha|^2 \sin^2 \tilde{\theta}) \sin^2 \theta = \frac{1}{2} + \left(\frac{1}{2} - P_{\text{jump}}\right) \cos 2\tilde{\theta} \cos 2\theta \tag{2.100}$$

The last equation is obtained by substituting $|\beta|^2 = P_{\text{jump}}$, $|\alpha|^2 = 1 - P_{\text{jump}}$. The above formula holds regardless of the adiabatic condition. The adiabatic solution can be obtained from Eq. (2100) by setting $P_{\text{jump}} = 0$. That is

$$\text{Adiabatic:} \qquad P(\nu_e \to \nu_e) = \sin^2 \theta + \cos^2 \tilde{\theta} \cos 2\theta \approx \sin^2 \theta \tag{2.101}$$

The last equality holds because, at the core, ν_e is almost $\tilde{\nu}_2$, which means $\tilde{\theta} \approx \pi/2$ from Eq. (2.92).

We can solve P_{jump} under the condition that the density changes linearly in the resonance region, which is given by [219, 220][23]

$$P_{\text{jump}} = \exp\left(-\frac{\pi}{2}\gamma_m\right) = \exp\left(-\frac{\pi \Delta m^2 \sin^2 2\theta}{4E \cos 2\theta (\ln n_e/dx)}\right) \tag{2.102}$$

This formula can be obtained by applying the Landau-Zener formula of the level crossing [224, 225]. The angle $\tilde{\theta}$ is the mixing angle in matter produced at the center of the Sun (to be exact, the point where ν_e is produced). As the density is high at the center and $\tilde{\nu}_2 \simeq \nu_e$, we can put $\tilde{\theta} \approx \pi/2$. Then

$$P(\nu_e \to \nu_e) \approx \sin^2 \theta + P_{\text{jump}} \cos 2\theta \tag{2.103}$$

We show the results of numerical calculations for two parameter settings ($\tan^2 \theta = 0.002, 0.48$) as a function of $E_\nu/\Delta m^2$ in Figure 2.29.[24] Analytical solutions based on Eq. (2100) give almost the same result.

Triangle of the matter oscillation: Considering the range of the energy spectrum of the Sun (0 – 14 MeV), and referring to conditions Eqs. (2.85) and (2.91), we know that the allowed region for the matter oscillation is constrained inside a triangle [226]. The solutions corresponding to the three side boundaries are as follows:

(I) Boundary I: the adiabatic solution: $\Delta m^2 \approx 10^{-4}\,(\text{eV})^2$, $\theta^2 \gtrsim 10^{-4}$. This defines the upper side of the triangle in Figure 2.30 denoted as (I).

(II) Boundary II: Nonadiabatic solution: from $\gamma_m \simeq 1$

$$\left(\frac{\Delta m^2}{1\,\text{eV}^2}\right) \sin 2\theta \tan 2\theta = \text{constant} \approx 5 \times 10^{-8} \left(\frac{E}{10\,\text{MeV}}\right) \tag{2.104}$$

This defines the lower left side of the triangle in Figure 2.30, denoted as (II).

(III) Solution 3: Quasi-vacuum solution: $\sin^2 2\theta \approx 1$, $10^{-10} \lesssim \Delta m^2 \lesssim 10^{-4}$. This is derived from Eq. (2.100), which shows that $P(\nu_e \to \nu_e) = 1/2$ at $\sin^2 2\theta = 1$, independent of Δm^2 and energy. This defines the right side of the triangle with

---

[23] An exact solution for the exponential function has also been obtained [221, 222, 223].

[24] They correspond to small mixing angle (SMA) and large mixing angle (LMA) solutions to be described later in the paragraph "Solutions to the solar neutrino oscillations". They represent the same probability as Figure 2.30, projected on to the $E_\nu/\Delta m^2$ axis at fixed $\tan^2 \theta$.

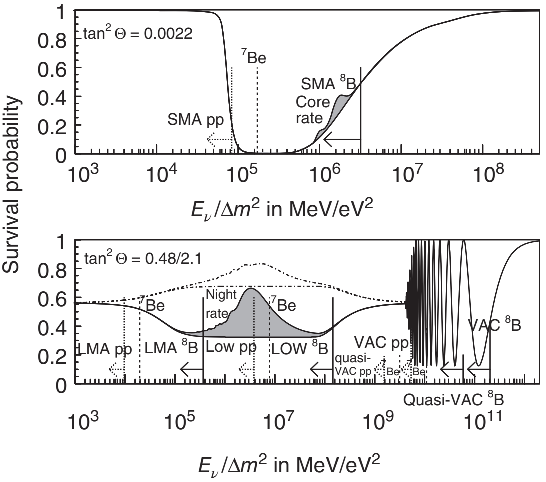


Figure 2.29 Survival probability for ^{8}B solar neutrinos for two fixed values of $\tan^2\theta$. The shaded area contrasts the difference between neutrinos passing through the Earth with neutrinos arriving directly from the Sun. Superimposed on these two panels are the locations of three neutrino branches (pp-dotted, 7Be-dashed, and 8B-solid) for the SMA, LMA, LOW, VAC, and quasi-VAC solutions (see Figure 2.31b for the abbreviated terminology). The line above the shaded area of the bottom figure depicts the average probability for neutrinos traversing Earth. (Reproduced with permission of [195].)

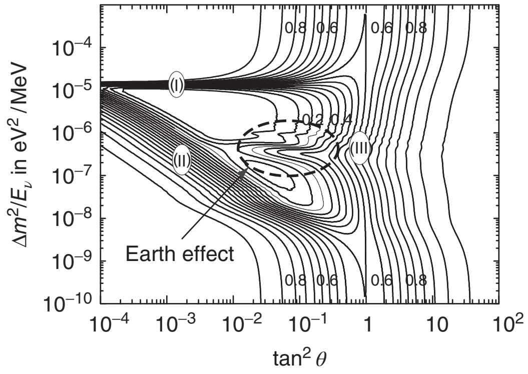


Figure 2.30 Contours of equal survival probability for ^{8}B solar neutrinos as a function of $\Delta m^2/E$ and $\tan^2\theta$. For the region (I), (II), and (III) see text. (Reproduced with permission of [195].)

additional allowed band stretching up and down from the side. It is the region denoted as (III) in Figure 2.30. In the extreme case, when Δm^2 is very small, the resonance condition is satisfied only close to the surface of the Sun. When the oscillation length is longer than the distance between the resonance position and the surface, the neutrinos do not fall into mass eigenstates, and the vacuum oscillation occurs at $\Delta m^2 \sim 10^{-(10-12)}$ eV^2.

Inside the triangle, the neutrino flux is depleted and the complete conversion occurs in some part of the area.

Earth effect: The matter effect exists also for the neutrino going through the Earth which recovers some of the loss in the Sun. This could be observed as the day-night difference. Inclusion of the earth effect modifies the contour of the quasi-vacuum boundary which sticks out to a smaller mixing angle region, as denoted in Figure 2.30.

Sharp descent to the valley of the vanishing probability in the upper panel of Figure 2.29 corresponds to the adiabatic solution, and the slow-rising slope on the right corresponds to the extreme nonadiabatic solution. The earth effect adds a bump to the curves. Also, the vertical lines correspond to three branches of pp, ^{7}Be, and ^{8}B neutrinos.

Solution to the Solar Neutrino Oscillation We have measurements on the total flux by chlorine, gallium, as well as SuperK and SNO data. The SuperK and SNO have additional information on the energy spectrum and day-night asymmetry.

Total flux: The total flux obtained by each group is shown in Figure 2.31a. For a given total flux at fixed energy, one can determine a triangle plus quasi-vacuum band sticking up and down from the right side of the triangle.

Since the energy regions that are measured by chlorine, SuperK, and SNO are similar, their triangles almost overlap. Their triangles and that by the Ga experiment are drawn on the $\tan^2\theta - \Delta m^2/E_\nu$ plane (Figure 2.31b), which constrains the

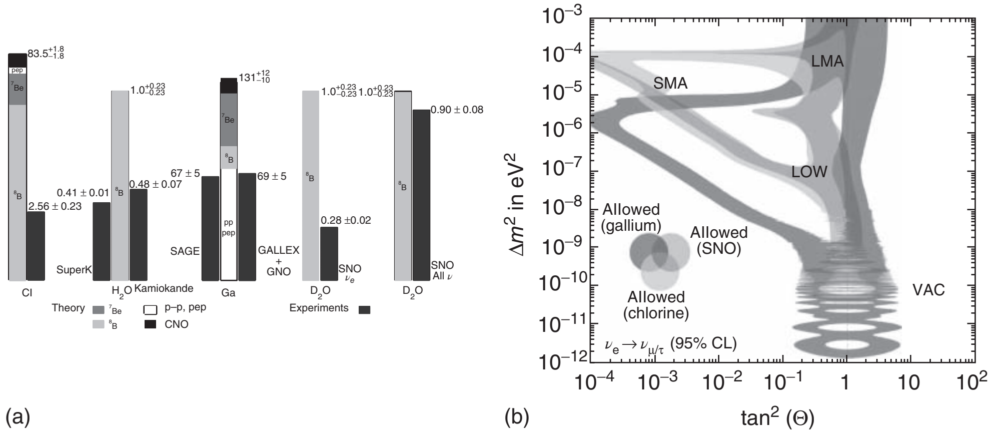


Figure 2.31 (a) Black histograms showing the observed total flux on the Cl, H_2O, Ga, D_2O target detectors. The middle/left histogram of each target shows the theoretical predictions for ^{8}B, ^{7}Be neutrinos with no oscillation. (Reproduced with permission of [227, 207].) (b) Allowed regions obtained from total flux rates. The "Cl/SNO" and "Ga" total fluxes constitute two triangular bands. "H_2O" and "D_2O" have extra information of energy spectra and day-night effects. Four solutions (LMA, SMA, LOW, VAC) remain as overlapped regions of all data. (Reproduced with permission of [228].)

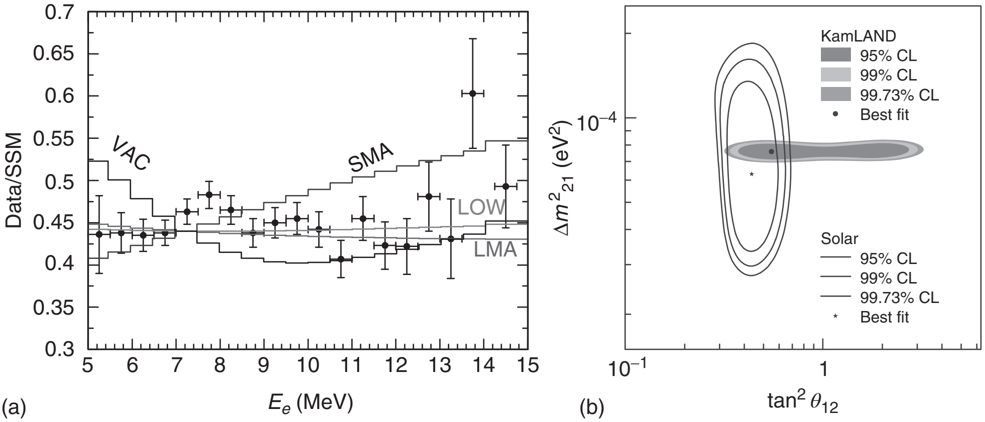


Figure 2.32 (a) Electron recoil energy spectrum measured in SuperK normalized to the SSM prediction, and the expectations for the best-fit points for the LMA, SMA, LOW, and VAC solutions in Figure 2.31b. [229, 195]. (b) Allowed region for neutrino oscillation parameters from KamLAND and solar neutrino experiments. (Reproduced with permission of [230].)

allowed region of the solar oscillation to four regions denoted as LMA, SMA, LOW, and VAC (sometimes referred as $JUSTSO$).

Energy spectrum and day-night effect: The energy spectra of the four solutions are given in Figure 2.32a.

The energy spectra of the SuperK and SNO made additional constraints. Both groups did not observe the day-night asymmetry, that is, they saw no earth effects. Using all the solar data, the solutions are reduced to LMA and LOW. The definitive data to determine the final solution to be LMA came from a reactor experiment by KamLAND. KamLAND observed a deficit in the reactor neutrino, which is $\overline{\nu}_e$. Using the CPT theorem, the $\overline{\nu}_e$ disappearance rate is the same as that of ν_e. The E/L of their data had the right value to observe the LMA solution. Their data, together with all the solar data combined results, are shown in Figure 2.32b. Later observation of ^{7}Be neutrino by Borexino supports the conclusion (see Figure 2.33).

The combined results give [8]

$$\Delta m^2_{21} = 7.50^{+0.19}_{-0.20} \times 10^{-5}\ \text{eV}^2 \tag{2.105a}$$

$$\sin 2\theta_{12} = 0.857^{+0.023}_{-0.025} \tag{2.105b}$$

Notice that, unlike the atmospheric neutrino result, there is no sign ambiguity of Δm^2_{21}. This is because the matter oscillation introduces asymmetry of the $\sin^2\ 2\theta$ with respect to $\pi/4$, as is clear from the use of $\tan^2\theta$ rather than $\sin^2\ 2\theta$.

Oscillation curve: Because of its large size, KamLAND had ability to detect reactor neutrinos from distant sources as far as 1000 km apart. Not all the reactors are in operation at the same time and, depending on the operation schedule, the effective E/L changes. Figure 2.34 shows the measured $\overline{\nu}_e$ flux as a function of E/L. This is an oscillation curve with $\Delta m^2 = 7.6 \times 10^{-5}$, which is relevant to the solar neutrino

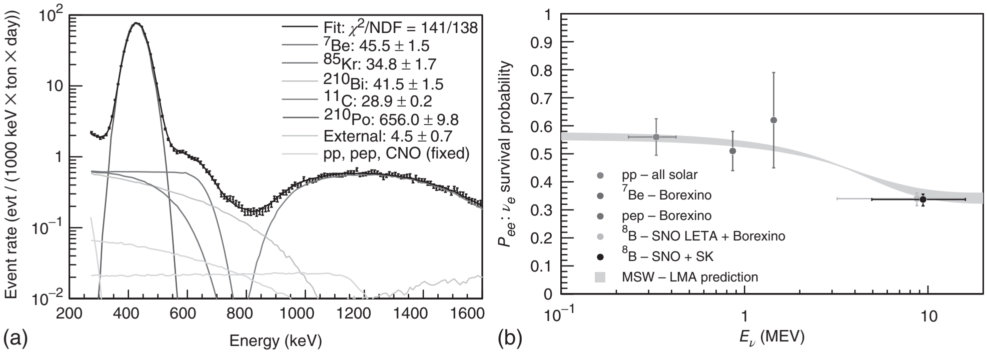


Figure 2.33 Borexino results on ^{7}Be solar neutrino observation. (a) Event rate as a function of the scattered electron energy in $\nu_e + e^- \rightarrow \nu_e + e^-$ reaction in the energy region 270–1600 keV. The solid line shows a Monte Carlo-based fitted spectrum (counts day^{-1} 100 ton^{-1})]. Some, but not all, of the α events have been removed using a cut. (b) Electron neutrino survival probability as a function of energy. The red line and ^{7}Be measurements of Pee are obtained in this experiment. The pp and ^{8}B measurements are also shown. The MSW–LMA prediction is shown for comparison; the band defines the 1σ range of the mixing parameter estimate in [7]. (Reproduced with permission of [231, 232].) (Please find a color version of this figure on the color plates.)

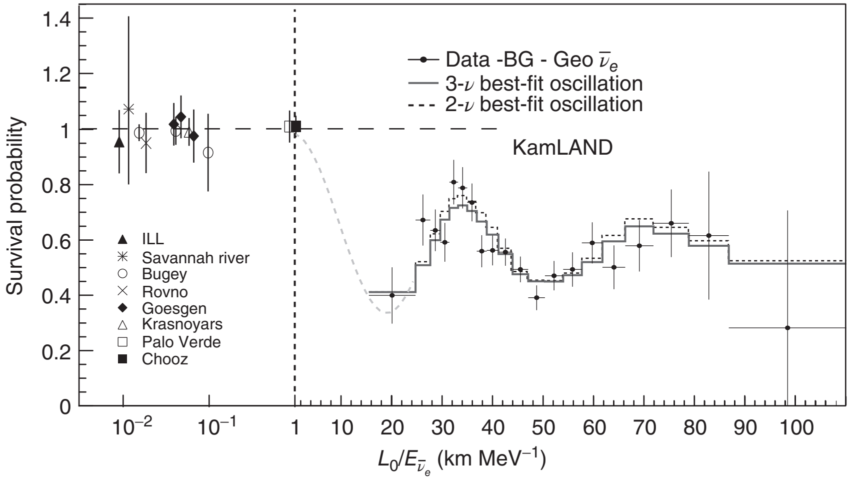


Figure 2.34 Ratio of measured to expected $\bar{\nu}_e$ flux from reactor experiments as a function of L_0/E. L_0 is the effective baseline taken as a flux-weighted average ($L_0 = 180$ km). The histogram and curve show the expectation accounting for the distances to the individual reactors, time-dependent flux variations, and efficiencies. [201, 233, 230] (Figure courtesy of [234].)

but realized on the terrestrial ground [compare it with Figure 2.9 for atmospheric neutrino]. The KamLAND data covers almost two cycles of the oscillation. Also shown are previous reactor results at much smaller distances. We can say that we now have solid evidence of the neutrino oscillation.

2.7.3
Reactor Experiment

In view of the important role the nuclear reactor played in the solar and anticipated future oscillation investigation, we will briefly describe the know-how of the reactor experiments as exemplified by KamLAND.

KamLAND: The KamLAND detector/target consists of 1 kt of ultrapure LS (liquid scintillator) contained in a 13-m-diameter spherical balloon made of a 135-μm-thick transparent nylon film (Figure 2.35). A buffer of mineral oil (MO) between the balloon and a 18-m-diameter spherical stainless steel containment vessel shields the LS from external radiation. The whole apparatus is placed in the old Kamiokande detector site filled with water (OD). There are 1325 newly developed, fast, 17-in.-diameter photomultipliers (PMTs) placed at the inner surface of the container and five-hundred and fifty-four 20-in. PMTs at the outer wall. The total photocathode coverage is 34%. A 3-mm-thick acrylic barrier of 16.6 m diameter reduces radon from PMT glass in the LS. The OD absorbs γ- rays and neutrons from the surrounding rock and acts as a tag for cosmic-ray muons. The energy resolution is $6.2\%/\sqrt{E\,(\mathrm{MeV})}$, and position determination accuracy is ~25 cm. The scintillator is purified to reduce the background. The residual isotope contents are given in Table 2.3. It is a trillion times cleaner than ordinary material (~1 ppm) and 100 times cleaner than SuperK. For reference, additional requirements are also listed in case detection of ^{7}Be solar neutrinos are to be carried out [201].

Energy spectrum of the neutrino: The neutrinos are produced in the β decays of the fission and decay products of the various isotopes (^{235}U, ^{238}U, ^{239}Pu, and ^{241}Pu). Figure 2.36a shows the energy spectra of the produced $\overline{\nu}_e$'s.

The observed yields can be obtained by multiplying the flux with the reaction cross section, and is depicted in Figure 2.36b. The observed signals in the experiment

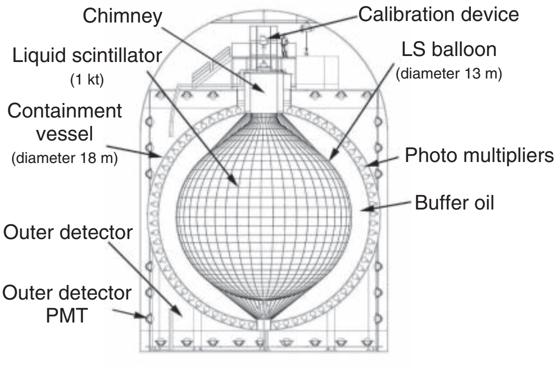


Figure 2.35 Schematic diagram of the KamLAND detector. (Reproduced with permission of [235, 236].)

Table 2.3 Requirements and achievements at KamLAND of radioactive impurities [201].

Impurities	Achievements	Required (reactor)	Required (solar ^{7}Be)
^{238}U	$3.5 \pm 0.5 \times 10^{-18}$g g^{-1}	10^{-13}g g^{-1}	10^{-16}g g^{-1}
^{232}Th	$5.2 \pm 0.8 \times 10^{-17}$g g^{-1}	10^{-13}g g^{-1}	10^{-16}g g^{-1}
^{40} K n	$< 2.7 \times 10^{-16}$g g^{-1}	10^{-14}g g^{-1}	10^{-18}g g^{-1}
^{222}Rn	0.03 μBq m^{-3}		1 μBq m^{-3}
^{85} Kr	0.7 Bq m^{-3}		1 μBq m^{-3}
^{210}Pb	20 mBq m^{-3}		1 μBq m^{-3}

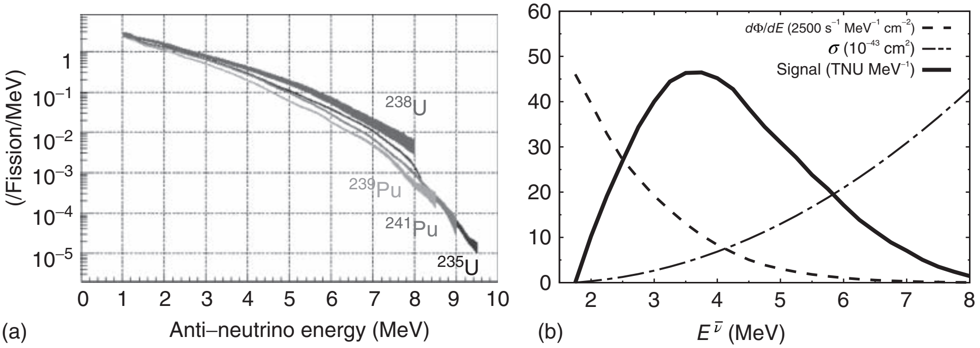


Figure 2.36 (a) Energy spectra of $\bar{\nu}_e$ flux from fission decays of ^{235}U,^{238} U,^{239} Pu, and^{241}Pu. [237, 238, 239, 240] The sum is known to reproduce the observed reactor flux to an accuracy of 1%. [241, 242] (b) Anti-neutrinos from reactors. The differential produced flux (no oscillations) (in units of 2500 $\bar{\nu}_e$ s^{-1}cm^{-1}MeV^{-1}), inverse beta decay cross section (in units of 10^{43}cm^2), and the corresponding signal (in TNU MeV^{-1}).[25] The flux corresponds to 10^{20} fissions s^{-1} (or about 3 GW) in a reactor 100 km from the detector. [243, 244] The detection efficiency is obtained by multiplying the $\bar{\nu}_e$ flux by the cross section, which results in a peak around 3 − 4 MeV. (Reproduced with permission of [245].)

are as follows:

$$\begin{aligned} &\bar{\nu}_e + p \to e^+ + n && E_{th} = \frac{(m_n + m_e)^2 - m_p^2}{2m_p} = 1.806\,\text{MeV} \\ &e^+ + e^- \to 2\gamma' s && E_{vis} \sim E_\nu - 0.78\,\text{MeV} > 1.022\,\text{MeV} \\ &n + p \to d + \gamma && \text{Delayed signal } (\tau{\sim}210\,\mu s), \quad E_\gamma = 2.22\,\text{MeV} \end{aligned} \tag{2.106}$$

Note that the energy spectrum below $E_p < 2.6$ MeV includes geo-neutrinos from the Earth's crust, and $E_{\text{cut}} = 2.6$ MeV is applied to select the reactor neutrinos.

---

[25] Terrestrial neutrino unit = one event/10^{32} target nuclei per year or 3.17×10^{-40}s^{-1} per target.

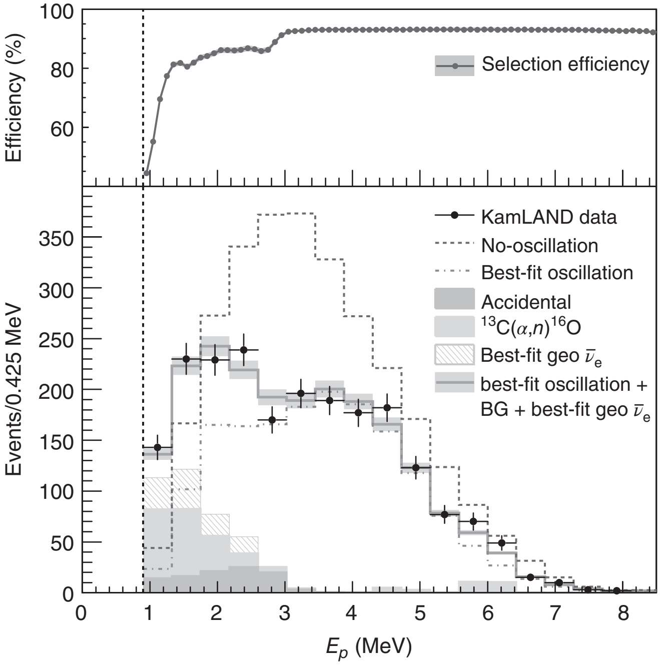


Figure 2.37 Prompt energy spectrum of $\bar{\nu}_e$ candidate events above 0.9 MeV energy threshold (vertical dashed line). The data together with the background and reactor $\bar{\nu}_e$ contributions fitted with oscillation hypothesis are shown in the main panel. The top panel shows the energy-dependent selection efficiency. The shaded background histograms are cumulative. 16 TW of earth's heat will produce nearly nine events (0.9 MeV) and nearly 0.04 events (2.6 MeV). The number of geo$\bar{\nu}_e$'s is unconstrained in the fit. (Reproduced with permission of [187].)

Figure 2.37 shows the observed energy spectrum of $\bar{\nu}_e$ obtained by KamLAND [187].

The observed data show a clear reduction compared to the no-oscillation hypothesis. Also shown are geo-neutrinos, which offer a new and useful method in clarifying the geo-heat source.

Backgrounds: The major backgrounds are several. The most serious source of external γ-rays from ^{208}Tl($E_\gamma \leq 3$ MeV) is strongly suppressed by the fiducial volume cut. At higher energies, the background is dominated by spallation products from energetic muons. Single neutrons are easily suppressed by 2-ms vetos following the muons. The dominant irreducible background is caused by $^{13}C(\alpha, n)^{16}O$ reactions from α-decay of ^{210}Po, a daughter of ^{222}Rn introduced into the LS during construction. A ^{210}Po^{13}C source was employed to study the $^{13}C(\alpha, n)^{16}O$ reaction and estimate the extent of their effect [246, 247].

2.8
Three-Flavor Oscillation

2.8.1
PMNS Matrix

So far we have used formulas valid for oscillations between two flavors. Actually, we have to do three-flavor analysis to obtain the true values of the parameters. The mixing matrix in the lepton sector is referred to as the PMNS (Pontecorvo-Maki-Nakagawa-Sakata) matrix [163, 164], which corresponds to CKM (Cabibbo-Kobayashi-Maskawa) matrix in the quark sector. In the standard parameterization, it can be expressed as

$$\begin{bmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{bmatrix} = U_{\text{PMNS}} \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{bmatrix} \quad ^{26)} \tag{2.107}$$

$$\begin{aligned} U_{\text{PMNS}} &\equiv \begin{pmatrix} U_{e1} & U_{e2} & U_{e3} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{pmatrix} \begin{pmatrix} e^{i\frac{\alpha_1}{2}} & 0 & 0 \\ 0 & e^{i\frac{\alpha_2}{2}} & 0 \\ 0 & 0 & 1 \end{pmatrix} \equiv UV \\ U &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{bmatrix} \begin{bmatrix} c_{13} & 0 & s_{13}e^{-i\delta} \\ 0 & 1 & 0 \\ -s_{13}e^{i\delta} & 0 & c_{13} \end{bmatrix} \begin{bmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix} \end{aligned} \tag{2.109}$$

where c_{ij}, s_{ij} denote $\cos\theta_{ij}$, $\sin\theta_{ij}$. Although we use the same notation for mixing angles θ_{ij} and the CP-violating Dirac phase δ, they are generally different from those of CKM matrix. Another important difference is the extra Majorana phase factors $e^{i\alpha_1/2}$, $e^{i\alpha_2/2}$. This is to take into account the possibility that the neutrino is a Majorana field because, if it is, unlike the Dirac field, it has no freedom of independent phase transformation, and the phase cannot be absorbed in the neutrino field redefinition [248, 249]. Notice that the Majorana phases do not appear in the oscillation, and hence we will omit them in the following discussions. However, they are important in the double-beta decay analysis, which will be treated

---

26) Note ν_α $(\alpha = e, \mu, \tau)$ are not kets (state vectors) but wave functions defined by

$$\nu_\alpha(t) = <\nu_\alpha|\psi(t)>, \quad |\psi(t)>= \sum_\alpha |\nu_\alpha><\nu_\alpha|\psi(t)>, \quad i\frac{\partial}{\partial t}|\psi(t)>= H|\psi(t)> \tag{2.108}$$

See [1]-(10.26).

in the next section. The oscillation probability in vacuum is given by

$$\begin{aligned}P(\nu_\alpha \to \nu_\beta) &= |<\nu_\beta|\nu_\alpha(t)>|^2 = \left|\sum_j U^*_{\alpha j} U_{\beta j} e^{-iE_j t}\right|^2 \\ &= \sum_{j,k} U^*_{\alpha j} U_{\beta j} U_{\alpha k} U^*_{\beta k} e^{-i(E_j - E_k)t} \\ &= \sum_j |U_{\alpha j}|^2 |U_{\beta j}|^2 + \sum_{j \neq k} U^*_{\alpha j} U_{\beta j} U_{\alpha k} U^*_{\beta k} e^{-i(E_j - E_k)t}\end{aligned} \tag{2.110}$$

where $j = 1 - 3$ are indexes of the mass eigenstates. Substituting the unitarity condition

$$\delta_{\alpha\beta} = \left|\sum_j U^*_{\alpha j} U_{\beta j}\right|^2 = \sum_j |U_{\alpha j}|^2 |U_{\beta j}|^2 + \sum_{j \neq k} U^*_{\alpha j} U_{\beta j} U_{\alpha k} U^*_{\beta k} \tag{2.111}$$

into Eq. (2.110), we have

$$\begin{aligned}P(\nu_\alpha \to \nu_\beta) &= \delta_{\alpha\beta} + \sum_{j \neq k} U^*_{\alpha j} U_{\beta j} U_{\alpha k} U^*_{\beta k} \left(e^{-i(E_j - E_k)t} - 1\right) \\ &\simeq \delta_{\alpha\beta} - 4\sum_{j>k} \mathrm{Re}\left[U^*_{\alpha j} U_{\beta j} U_{\alpha k} U^*_{\beta k}\right] \sin^2 X_{jk} + 2\sum_{j>k} \mathrm{Im}\left[U^*_{\alpha j} U_{\beta j} U_{\alpha k} U^*_{\beta k}\right] \sin 2X_{jk}\end{aligned} \tag{2.112a}$$

$$X_{jk} \equiv \frac{m_j^2 - m_k^2}{4E} ct = 1.27 \frac{\Delta m^2_{jk}}{(\mathrm{eV}^2)} \frac{L/E}{(\mathrm{m/MeV})} \tag{2.112b}$$

where $\simeq$ is due to the relativistic approximation $E_j \simeq p + m_j^2/2p \simeq p + m_j^2/2E$. The formula for anti-neutrinos can be obtained simply by changing the sign of δ, that is

$$P(\overline{\nu}_\alpha \to \overline{\nu}_\alpha; \delta) = P(\nu_\alpha \to \nu_\beta; -\delta) \tag{2.113}$$

This follows from the CPT and T transformation relations:

$$P(\overline{\nu}_\alpha \to \overline{\nu}_\beta) \overset{\mathrm{CPT}}{=} P(\nu_\beta \to \nu_\alpha) \tag{2.114a}$$

$$P(\nu_\beta \to \nu_\alpha; \delta) \overset{\mathrm{T}}{=} P(\nu_\alpha \to \nu_\beta; -\delta) \tag{2.114b}$$

Thus, going from two to three flavors, we have obtained a CP-violating term. This is similar to the CKM matrix in the quark sector.

CP-Violating Term The formula for CP conjugate oscillation probability $P(\overline{\nu}_\alpha \to \overline{\nu}_\beta)$ can alternatively be obtained from Eq. (2.112) by replacing $U \to U^*$, which changes only the sign of the third term. Therefore, by taking the difference, we can extract the CP-violating terms:

$$\Delta^{\mathrm{CP}}_{\alpha\beta} \equiv \frac{1}{2}\left[P(\nu_\alpha \to \nu_\beta) - P(\overline{\nu}_\alpha \to \overline{\nu}_\beta)\right] = 2\sum_{j>k} \mathrm{Im}\left[U^*_{\alpha j} U_{\beta j} U_{\alpha k} U^*_{\beta k}\right] \sin 2X_{jk}, \ \alpha \neq \beta \tag{2.115}$$

Note that, for $\alpha = \beta$, no CP-violating term appears. The equation can be simplified by using the unitarity relation

$$U_{\alpha i}U^*_{\beta i} + U_{\alpha j}U^*_{\beta j} + U_{\alpha k}U^*_{\beta k} = 0 \qquad i, j, k = 1, 2, 3 \text{ and cyclic} \tag{2.116}$$

Multiplying $U^*_{\alpha k}U_{\beta k}$, one obtains

$$\begin{aligned}
&(U^*_{\alpha k}U_{\beta k}U_{\alpha i}U^*_{\beta i}) + (U^*_{\alpha k}U_{\beta k}U_{\alpha j}U^*_{\beta j}) + |U_{\alpha k}|^2|U_{\beta k}|^2 = 0\\
&\therefore \operatorname{Im}(U^*_{\alpha j}U_{\beta j}U_{\alpha k}U^*_{\beta k}) = \operatorname{Im}(U^*_{\alpha k}U_{\beta k}U_{\alpha i}U^*_{\beta i}) = \operatorname{Im}(U^*_{\alpha i}U_{\beta i}U_{\alpha j}U^*_{\beta j}) = \mp J\\
&J = \mp \operatorname{Im}(U^*_{\alpha k}U_{\beta k}U_{\alpha i}U^*_{\beta i})\,(\alpha \neq \beta,\ k \neq i) = -\operatorname{Im}\Big(U^*_{e3}U_{\mu 3}U_{e1}U^*_{\mu 1}\Big)\\
&\quad = c_{12}s_{12}c_{23}s_{23}c^2_{13}s_{13}\sin\delta = \frac{1}{8}c_{13}\sin 2\theta_{12}\sin 2\theta_{23}\sin 2\theta_{31}\sin\delta
\end{aligned} \tag{2.117}$$

where we have substituted the PMNS matrix elements of Eq. (2.109) explicitly. The third equality of the second line is obtained by cyclic permutation. The sign to be adopted is (−) for $(\alpha, \beta) = (e, \mu)$, (μ, τ), (τ, e), and $j, k = (1, 2)$, $(2, 3)$, $(3, 1)$. The expression J is referred to as the *Jarlskog factor* and represents twice the area of the unitary triangle defined by Eq. (2.116) in the complex plane. One sees that the mixing matrix part of Eq. (2.115) can be factored out. Furthermore, using the identity

$$\begin{aligned}
4\sin A\sin B\sin C = {}&\sin(A + B - C) + \sin(B + C - A) + \sin(C + A - B)\\
&- \sin(A + B + C)
\end{aligned} \tag{2.118}$$

and substituting $A = X_{12}$, $B = X_{23}$, $C = X_{31}$, $X_{12} + X_{23} + X_{31} = 0$ in the above equation, we have the relation

$$\sin 2X_{12} + \sin 2X_{23} + \sin 2X_{31} = -4\sin X_{12}\sin X_{23}\sin X_{31} \tag{2.119}$$

Substituting Eq. (2.119) in Eq. (2.115), the CP-violating term is expressed as

$$\begin{aligned}
&\Delta^{\text{CP}}_{\alpha\beta} \equiv \frac{1}{2}\left[P(\nu_\alpha \to \nu_\beta) - P(\overline{\nu}_\alpha \to \overline{\nu}_\beta)\right] = \begin{cases} -8J\sin X_{12}\sin X_{23}\sin X_{31} & \alpha \neq \beta\\ 0 & \alpha = \beta \end{cases}\\
&\qquad (\alpha,\ \beta) = (e, \mu),\ (\mu, \tau),\ \text{ or }\ (\tau, e)
\end{aligned} \tag{2.120}$$

One sees that the CP violation occurs only when all the mixing angles and mass differences do not vanish.

First order formulas: The oscillation probability depends on the mixing angles θ_{ij} and the mass squared differences $X_{ij} = \Delta m^2_{ij}(L/4E)$. Their exact expressions are rather lengthy, in general. In order to simplify the formula, we notice that Δm^2 for the atmospheric and solar neutrinos are different by more than an order of magnitude.

$$\alpha \equiv \frac{\Delta m^2_\odot}{|\Delta m^2_{\text{atm}}|} = \frac{\Delta m^2_{21}}{|\Delta m^2_{31}|} = \frac{7.7 \times 10^{-5}}{2.41 \times 10^{-3}} \simeq 0.032 \tag{2.121}$$

which means a big difference in the oscillation wavelength. Because of this, terms containing $X_\odot \equiv X_{21}$ can be neglected for terrestrial short-baseline (≪100 km) experiments. On the other hand, for the solar neutrino observation, the baseline is so long that the oscillation due to $X_{\text{atm}} = X_{31} = X_{32} + X_{21} \simeq X_{32}$ can be averaged ($<\sin^2 X_{\text{atm}}> \approx 1/2$). As a result, using Eq. (2.112a), the three-flavor oscillation probabilities in vacuum can be approximated as [250]

$$\begin{aligned}
P^{3\nu}(\nu_e \to \nu_e) &= 1 - c_{13}^4 \sin^2 2\theta_{12} \sin^2 X_{21} - \sin^2 2\theta_{13}\left[c_{12}^2 \sin^2 X_{31} + s_{12}^2 \sin^2 X_{32}\right] \\
&\simeq 1 - \sin^2 2\theta_{13} \sin^2 X_{\text{atm}} - c_{13}^4 \sin^2 2\theta_{12} \sin^2 X_{\odot} \qquad (2.122\text{a}) \\
P^{3\nu}(\nu_\mu \to \nu_\mu) &= 1 - P^{3\nu}(\nu_\mu \to \nu_e) - P^{3\nu}(\nu_\mu \to \nu_\tau) \\
&\simeq 1 - 4 s_{13}^2 c_{13}^2 s_{23}^2 \sin^2 X_{31} - c_{13}^4 \sin^2 2\theta_{23} \sin^2 X_{32} \\
&\simeq 1 - (c_{13}^4 \sin^2 2\theta_{23} + \sin^2 2\theta_{13} s_{23}^2) \sin^2 X_{32} \qquad (2.122\text{b}) \\
P^{3\nu}(\nu_e \to \nu_\mu) &\simeq \sin^2 2\theta_{13} s_{23}^2 \sin^2 X_{32} \qquad (2.122\text{c}) \\
P^{3\nu}(\nu_\mu \to \nu_\tau) &\simeq c_{13}^4 \sin^2 2\theta_{23} \sin^2 X_{32} = c_{13}^4 P^{2\nu}(\nu_\mu \to \nu_\tau) \qquad (2.122\text{d})
\end{aligned}$$

where $P^{2\nu}$ is the two-flavor oscillation formula. The CP-violating term does not appear in this approximation.

A caution in applying $P(\nu_e \to \nu_e)$ to the reactor neutrino: as the energy of the reactor neutrino is low (~4 MeV), the second term in Eq. (2.122a) reaches a maximum at ~2 km, and the third term at ~70 km (corresponding to L/E ~0.5 and 17, respectively) (see Figure 2.38).

The former applies to short-baseline experiments, which had difficulty in observing the oscillation effects [178, 179, 251] until recently. The latter applies to KamLAND for terrestrial observation of 1 − 2 (i.e., solar) oscillation, which indeed detected the $\bar{\nu}_e \to \bar{\nu}_x$ transition and confirmed the LMA solution of the solar neutrino. As one can see from Eq. (2.122a), $P^{3\nu}(\nu_e \to \nu_e)$ reduces to $P^{2\nu}(\nu_e \to \nu_e)$ [see

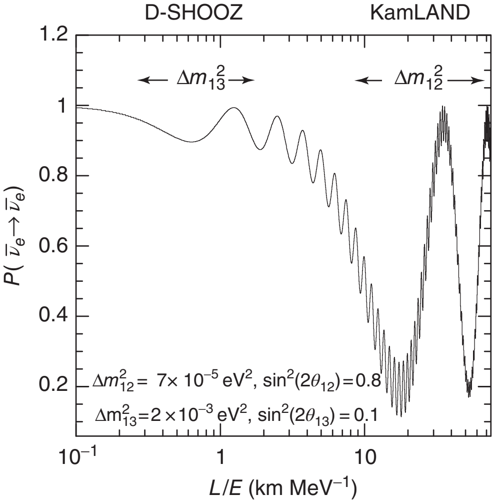


Figure 2.38 In the reactor neutrino oscillation $[P(\bar{\nu}_e \to \bar{\nu}_e)]$, the atmospheric and solar maxima appear at L/E ~0.5 and 17 (or $L \sim 1.4$ km and ~70 km for $E \sim 4$ MeV), respectively. The survival probability of $\bar{\nu}_e$ versus the ratio of the distance to the neutrino energy (L/E) is drawn with $\sin^2(2\theta_{13})$ set at the maximum value allowed by the past measurements.

Eqs. (2.49) and (2.51)] in the approximation $\sin^2\theta_{13} \simeq 0$. In the same approximation, $P^{3\nu}(\nu_\mu \to \nu_e) \simeq 0$ [Eqs. (2.122c) and (2.114)] and $P^{3\nu}(\nu_\mu \to \nu_\mu) \simeq P^{2\nu}(\nu_\mu \to \nu_\mu) \simeq 1 - P^{2\nu}(\nu_\mu \to \nu_\tau)$ [Eqs. (2.122b) and Eq. (2.122d)]. That is, within the approximation to set $\theta_{13} = 0$, we are justified to apply the two-flavor analysis to have a quick look of the oscillation parameters. However, the three-flavor analysis has to be used to obtain accurate values, in particular the mixing angle θ_{13} and the CP phase δ.

Mixing Angle θ_{13} Although global fits to all the available oscillation data [252, 253, 254] or later experiments [187] saw some hints of nonvanishing $\sin^2 2\theta_{13}$, conclusions were not definitive. Consequently, dedicated experiments to look for especially $\sin^2 2\theta_{13}$, that is, the short-baseline ($\lesssim 1 - 2$ km) reactor experiments, with sensitivity about 10 times better than previous ones were constructed [255, 256, 257]. The signal to look for was the disappearance of the $\overline{\nu}_e$. Using the CPT invariance and Eq. (2.122a), the transition probability is given by

$$P(\overline{\nu}_e \to \overline{\nu}_e) = P(\nu_e \to \nu_e) \simeq 1 - \sin^2\theta_{13}\sin^2 X_{13} \tag{2.123}$$

All the three experiments detected the disappearance of the flux and obtained a finite value for the mixing angle θ_{13}.

Here we describe the experiment carried out at Daya Bay in China [255]. Antineutrinos from six 2.9-GW_{th} reactors were detected in six anti-neutrino detectors (ADs) deployed in two near (flux-weighted baseline 470 m and 576 m) and one far (1648 m) underground experimental halls (Figure 2.39a). $\overline{\nu}_e$ was detected via the inverse beta decay reaction, $\overline{\nu}_e + p \to e^+ + n$, in a gadolinium-doped liquid scintillator (Gd-LS). The coincidence of the prompt scintillation from e^+ and the delayed neutron capture on Gd provides a distinctive $\overline{\nu}_e$ signature.

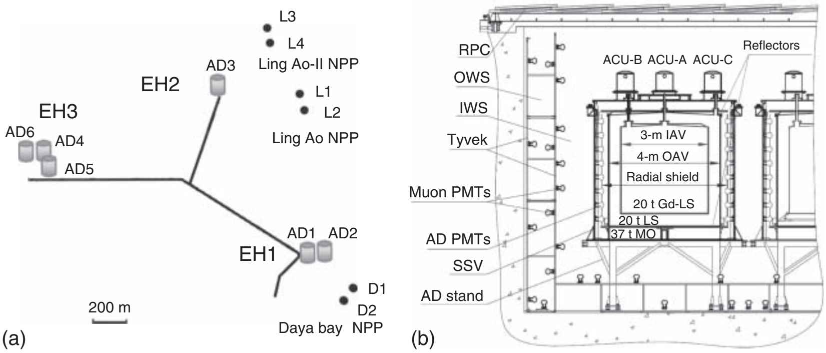


Figure 2.39 (a) Layout of the Daya Bay experiment. The dots represent reactors, labeled D1, D2, L1, L2, L3, and L4. Six identical ADs, AD1–AD6, are installed in three experimental halls (EH1–3). (b) Schematic diagram of the Daya Bay detectors. The target is a gadolinium-doped liquid scintillator (Gd-LS) surrounded by two layers of radiation shields [LS (liquid scintillator) and MO (mineral oil)]. They are further surrounded by inner and outer water shields (IWS, OWS). (Reproduced with permission of [255].)

Each AD consists of a cylindrical, 5-m-diameter, stainless steel vessel (SSV) that houses two nested, UV-transparent, acrylic cylindrical vessels (Figure 2.39b). A 3.1-m-diameter inner acrylic vessel (IAV) holds 20 t of Gd-LS (target). It is surrounded by a region with 20 t of LS inside a 4-m- diameter outer acrylic vessel (OAV). Between the SSV and OAV, 37 t of MO shields the LS and Gd-LS from radioactivity. They are shielded with >2.5 m of high-purity water against ambient radiation in all directions. Each water pool is segmented into inner water shield (IWS)and outer water shield (OWS) and instrumented with PMTs to function as Cherenkov-radiation detectors to remove spallation neutrons and other cosmogenic backgrounds.

With a 43,000-ton GW$_{th}$-day live-time exposure in 55 days, 10,416 (80,376) electron–antineutrino candidates were detected at the far hall (near halls). The ratio of the observed to expected number of anti-neutrinos at the far hall is $R = 0.940 \pm 0.011(\text{stat}) \pm 0.004(\text{syst})$. Figure 2.40b shows the reactor $\overline{\nu}_e$ rate as a function of distance. A rate-only analysis finds $\sin^2 2\theta_{13} = 0.092 \pm 0.016(\text{stat}) \pm 0.005(\text{syst})$ in the three-neutrino framework.

The combined average of the three experiments is given by [8]

$$\sin^2\ 2\theta_{13} = 0.098 \pm 0.013 \tag{2.124}$$

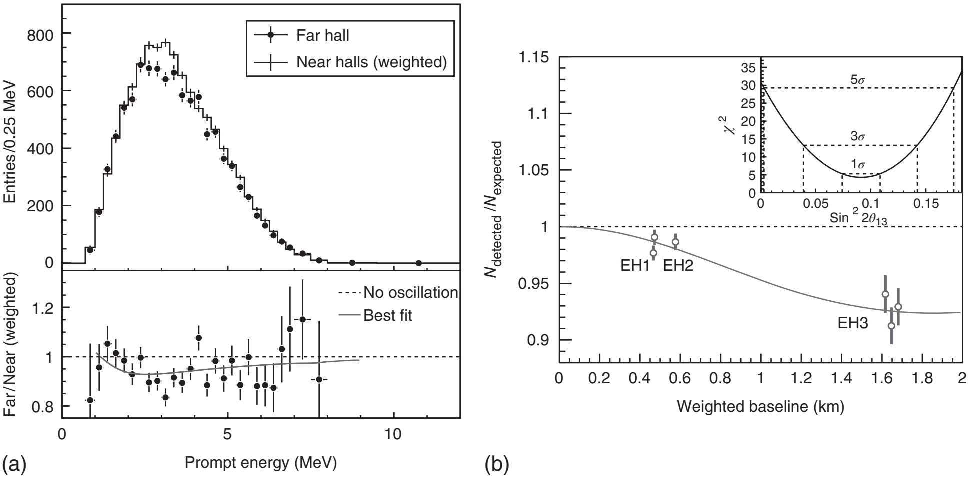


Figure 2.40 Daya Bay reactor experiment (a) Top: Measured prompt-energy spectrum of the far hall [sum of three AD (anti-neutrino detector)s] compared with the no-oscillation prediction from the measurements of the two near halls. Spectra were background-subtracted. Uncertainties are statistical only. Bottom: Ratio of the measured and predicted no-oscillation spectra. The solid curve is the best-fit solution with $\sin^2 2\theta_{13} = 0.092$ obtained from the rate-only analysis. The dashed line is the no-oscillation prediction. (b) Ratio of measured versus expected signal in each detector, assuming no oscillation. The error bar is the uncorrelated uncertainty of each AD, including statistical, detector-related, and background-related uncertainties. Reactor and survey data were used to compute the flux-weighted average baselines. The oscillation survival probability at the best-fit value is given by the smooth curve. The AD4 and AD6 data points are displaced by −30 and +30 m for visual clarity. The χ^2 versus $\sin^2 2\theta_{13}$ is shown in the inset. (Reproduced with permission of [255].)

2.8.2
Summary of Experimental Data

We summarize current values of the oscillation parameters [8].

$$\left\{\begin{array}{llll} \theta_A & = \theta_{\text{atm}} = \theta_{23} \simeq \frac{\pi}{4}, & \sin^2 2\theta_{23} & > 0.95 \\ \Delta m^2_{\text{atm}} & = |\Delta m^2_{31}| \simeq |\Delta m^2_{32}| & & \simeq 2.32^{+0.12}_{-0.06} \times 10^{-3}\ \text{eV}^2 \\ \theta_{\odot} & = \theta_{12} = 32.5^{+2.4^{\circ}}_{-2.3} & \sin^2 2\theta_{12} & = 0.857^{+0.023}_{-0.025} \\ \Delta m^2_{\odot} & = \Delta m^2_{21} & & = 7.50^{+0.19}_{-0.20} \times 10^{-5}\ \text{eV}^2 \\ & & \sin^2 2\theta_{13} & = 0.098 \pm 0.013 \end{array}\right. \tag{2.125a}$$

$$U_{\text{PMNS}} \simeq \begin{bmatrix} c_{12} & s_{12} & s_{13}e^{-i\delta} \\ -\dfrac{s_{12}}{\sqrt{2}} & \dfrac{c_{12}}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \\ \dfrac{s_{12}}{\sqrt{2}} & -\dfrac{c_{12}}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \end{bmatrix} \tag{2.125b}$$

One can see that U_{PMNS} is very different from the CKM matrix in the quark sector, which is nearly diagonal. One also notices that $\theta_{23} \simeq \pi/4$, $\sin\theta_{12} \simeq 1/\sqrt{3}$, which are very suspicious numbers. The PMNS matrix given by setting $\theta_{23} = \pi/4$, $\theta_{13} \simeq 0$, $\sin\theta_{12} = 1/\sqrt{3}$ is referred to as the *tri-bimaximal matrix* and $\theta_{12} = \pi/4$ as the *bimaximal matrix.* [27]

$$U_{tri\text{-}bi} \equiv \begin{bmatrix} \sqrt{\frac{2}{3}} & \frac{1}{\sqrt{3}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \end{bmatrix} \tag{2.126}$$

This means

$$\begin{aligned} |\nu_1> &= (2|\nu_e> - |\nu_\mu> + |\nu_\tau>)/\sqrt{6} \\ |\nu_2> &= (|\nu_e> + |\nu_\mu> - |\nu_\tau>)/\sqrt{3} \\ |\nu_3> &= (|\nu_\mu> + |\nu_\tau>)/\sqrt{2} \end{aligned} \tag{2.127}$$

The three neutrino mass eigenstates are very close to those realized by the tri-bimaximal matrix. Its origin is yet unknown [258, 259, 260, 261].

Mass Hierarchy of the Neutrino Of the two mass differences so far measured, $\Delta m^2_{\odot}$, is known to be positive ($\Delta m^2_{\odot} > 0$), that is, $m_2 > m_1$, but only the absolute value is known for Δm^2_{32}. Depending on its sign, there are two possibilities: one $m_3 > m_2 > m_1$ called *normal hierarchy* (*NH*), and the other $m_2 > m_1 > m_3$ called *inverted hierarchy* (*IH*). Absolute values of the neutrino mass cannot be determined from the oscillation experiments alone, but if one assumes hierarchical structure $m_3 \gg m_2 \gg m_1$, one obtains $m_3 \simeq \sqrt{|\Delta m^2_{32}|} \simeq 0.05$ eV and $m_2 \simeq \sqrt{|\Delta m^2_{21}|} \sim 0.008$

---

[27] The name "tri-bimaximal" comes from the fact that ν_2 and ν_3 are maximal mixtures of three- and two-flavor eigenstates as given in Eq. (2.127).

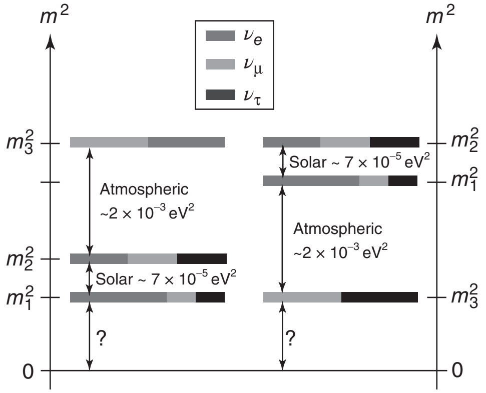


Figure 2.41 Level structure of the neutrino. The left column shows the normal hierarchy and the right one shows the inverted hierarchy. There is also a possibility of degenerate masses ($m_1 \simeq m_2 \simeq m_3 >> \sqrt{|\Delta m^2|}$), which will be described in Figure 2.45. The bands showing levels of the mass are color-coded by mixing ratios of the flavor states obtained from Eq. (2.125). (Please find a color version of this figure on the color plates.)

eV for NH (Figure 2.41). The level structure of the IH is also shown. In addition, there is a possibility that the three neutrinos possess relatively large masses but their differences are small (quasi-degenerate). However, a constraint from cosmological arguments [119] gives $\sum_j m_j \lesssim 0.14$ eV (see Figure 2.1), and is marginally larger than 0.05 eV, which is obtained by atmospheric neutrino observations.

Two unknown parameters; mass hierarchy; and CP parameter δ. Within the three-flavor framework, we still have two undetermined parameters, namely sign of Δm_{31}^2 and the CP phase δ. In view of the seesaw mechanism [262, 263] [see also discussions after Eq. (2.17) and leptogenesis (see Section 8.6.3)] being the primary candidate mechanism for the origin of the neutrino mass and cosmic baryon asymmetry, it is of paramount importance to clarify the origin of the CP violation and mass hierarchy. Determining all the mixing angles and CP parameters with accuracy comparable to that obtained for the CKM matrix is our next goal in neutrino physics. Current efforts on the accelerator experiments are mainly to improve the accuracy of $\nu_\mu \rightarrow \nu_e$ transition and determine those parameters (i.e., θ_{13}, sign of Δm_{atm}^2, and δ_{CP}). Two experiments, T2K at KEK [264] and NOvA [265], are ongoing and scheduled to run in 2013.[28]

2.8.3
CP Violation and Mass Hierarchy

Note that the short-baseline reactor experiments are capable of determining the magnitude of the $\sin^2 2\theta_{13}$ but has no sensitivity to the sign of Δm^2 or the CP phase. To determine them, long-baseline accelerator experiments are needed. However, as the length of the baseline becomes large ($L \gtrsim 300$ km), one can no longer neglect the earth matter effect. The Schrödinger equation governing the three-flavor oscillation is given by [267, 226]

---

[28] T2K announced a value of $\sin^2\theta_{13} = 0.08^{+0.049}_{-0.039}$ (stat + syst) in the $\nu_\mu \rightarrow \nu_e$ conversion process [266].

$$\begin{aligned} i\frac{d\nu_\alpha}{dL} &= \frac{1}{2E_\nu}\sum_\beta \left(A\delta_{\alpha e}\delta_{\beta e} + \sum_i U^*_{\beta i}U_{\alpha i}m_i^2 \right)\nu_\beta \\ &= \frac{1}{2E_\nu}\sum_\beta \left(A\delta_{\alpha e}\delta_{\beta e} + \sum_i U^*_{\beta i}U_{\alpha i}\Delta m_{i1}^2 \right)\nu_\beta \\ A &= 2\sqrt{2}G_F n_e E_\nu = 1.52\times 10^{-4}\ \text{eV}^2\ \rho\,(\text{g cm}^{-3})E\,(\text{GeV}) \end{aligned} \tag{2.128}$$

where G_F is the Fermi coupling constant of the weak interaction, $n_e = \rho N_A Z/A$ is the electron number density, $N_A = 6.022\times 10^{23}$ is the Avogadro number, and ρ is the density of matter. We used the unitary condition in deriving the second line of Eq. (2.128). For $L < 5000$ km, the earth density can safely be regarded as a constant. Treating the two parameters

$$\alpha \equiv \frac{\Delta m_{21}^2}{\Delta m_{31}^2} \simeq 0.032, \qquad \sin^2\theta_{13} \simeq 0.025 \tag{2.129}$$

as small numbers, and up to second order in these parameters, convenient and generic oscillation probabilities for the terrestrial experiments including the matter effect are given by [268, 269, 270]:

$$\begin{aligned} P(\nu_\mu \to \nu_e) &= s_{23}^2\sin^2 2\theta_{13}\frac{\sin^2[(1-\beta)X_{31}]}{(1-\beta)^2} && :\ \theta_{13}\ \text{term} \\ &\quad + \alpha\tilde{J}\cos(\delta_{\text{CP}} - X_{31})\frac{\sin(\beta X_{31})\sin[(1-\beta)X_{31}]}{\beta(1-\beta)} && :\ \text{CP term} \\ &\quad + \alpha^2 c_{23}^2\sin^2 2\theta_{12}\frac{\sin^2(\beta X_{31})}{\beta^2} && :\ \text{Solar term} \end{aligned} \tag{2.130}$$

$$P(\nu_\mu \to \nu_\mu) = 1 - c_{13}^2\sin^2 2\theta_{23}\sin^2 X_{31} + 2\alpha c_{13}^2 c_{12}^2\sin^2 2\theta_{23}\cos X_{31} \tag{2.131a}$$

$$P(\nu_e \to \nu_e) = 1 - \sin^2 2\theta_{13}\sin^2 X_{31} - \alpha^2 X_{31}^2 c_{13}^4\sin^2 2\theta_{12} \tag{2.131b}$$

where

$$X_{31} = \frac{\Delta m_{31}^2 L}{4E} = \frac{m_3^2 - m_1^2}{4E}L = \pm 3.05\times 10^{-3}\left(\frac{\Delta m_{31}^2}{2.41\times 10^{-3}\ \text{eV}^2}\right)\frac{L(\text{km})}{E(\text{GeV})} \tag{2.132a}$$

$$\begin{aligned} \beta = \frac{2\sqrt{2}G_F n_e E_\nu}{\Delta m_{31}^2} &= \left(\frac{E}{12\,\text{GeV}}\right)\left(\frac{\Delta m_{31}^2}{2.5\times 10^{-3}\ \text{eV}^2}\right)^{-1}\left(\frac{Y_e\,\rho}{1.4\,\text{g cm}^{-3}}\right) \\ &= \left(\frac{L}{5820\,\text{km}}\right)\left(\frac{\pi/2}{X_{31}}\right)\left(\frac{Y_e\,\rho}{1.4\,\text{g cm}^{-3}}\right) \end{aligned} \tag{2.132b}$$

$$\tilde{J} = 8J/\sin\delta = c_{13}\sin 2\theta_{12}\sin 2\theta_{23}\sin 2\theta_{31} \tag{2.132c}$$

where Y_e is the average Z/A. The matter effect is contained in β. The sign of X_{31} is (+) for normal and (−) for IH. $P(\overline{\nu}_\mu \to \overline{\nu}_e)$ can be obtained from $P(\nu_\mu \to \nu_e)$ by replacing $\delta_{\mathrm{CP}} \to -\delta_{\mathrm{CP}}$, $\beta \to -\beta$ and $P(\nu_e \to \nu_\mu)$ by replacing $\delta_{\mathrm{CP}} \to -\delta_{\mathrm{CP}}$ [271]. Note that the matter effect violates the CPT invariance. The above formulas are valid for $L < 5000$ km and $E \gtrsim 0.5$ GeV. For experiments with $L \lesssim 1000$ km, which apply both to T2K and NOvA (to be described soon), the matter effect is small and can be treated as a small perturbation. Also, at or near the oscillation maximum (i.e., at $\sin^2 X_{31} \approx 1$), which one may assume as an optimum experimental setting, Eq. (2.130) become up to $O(\beta)$, $O(\alpha)$

$$\begin{aligned}
&P(\nu_\mu \to \nu_e)\\
&\quad = s_{23}^2 \sin^2 2\theta_{13} \sin^2 X_{31}(1+2\beta) + \alpha X_{31}\tilde{J}\cos(\delta_{\mathrm{CP}} - X_{31}) \sin X_{31}\\
&\quad = s_{23}^2 \sin^2 2\theta_{13} \sin^2 X_{31}(1+2\beta)\\
&\qquad + \alpha X_{31}\tilde{J} \sin X_{31}\left\{ \cos X_{31} \cos\delta_{\mathrm{CP}} + \sin X_{31} \sin\delta_{\mathrm{CP}} \right\}
\end{aligned} \tag{2.133}$$

where we dropped the small α^2 term. Note that the CP-violating term is the $\sin\delta$ part of the second term. It agrees with Eq. (2.120) for $\alpha = \mu$, $\beta = e$ if we replace $\sin X_{21} \to \alpha X_{31}$ and approximate $X_{32} \approx X_{31}$. In the above expression, the first term is the dominant one. Now that the angle θ_{13} has been determined by the reactor experiments, one can proceed to look for the CP effect in the second term. The CP asymmetry is given by

$$A_{\mathrm{CP}} = \frac{P(\nu_\mu \to \nu_e) - P(\overline{\nu}_\mu \to \overline{\nu}_e)}{P(\nu_\mu \to \nu_e) + P(\overline{\nu}_\mu \to \overline{\nu}_e)} \simeq \frac{\Delta m_{21}^2 L}{4E_\nu} \frac{\sin 2\theta_{12}}{s_{13}} \sin\delta_{\mathrm{CP}} \tag{2.134}$$

We see that, to determine the CP phase, knowledge of an accurate value of θ_{13} is vital.

T2K and NOvA: In order to determine the CP term, a conventional detector arrangement is to have a first detector near the accelerator to determine the initial flux and a second detector at the far site. T2K [272, 264] and NOvA [265, 273] are second-generation experiments of K2K and MINOS, which were described in Section 2.5.3. An optimal condition is to set the far detector at the first oscillation maximum, that is, $X_{31} = \Delta m_{31}^2 L/4E = \pi/2$, which determines $L(\mathrm{km})/E(\mathrm{GeV}) = 515$. For $L = 295/735$ km (T2K/NOvA), it occurs at $E_\nu \sim 0.57/1.43$ GeV. T2K uses the SuperK detector as the far detector, while NOvA builds a newly designed fine-grained (385k cells) 15-kt LS calorimeter.

Both experiments expect to have the resolution $\delta(\sin^2 2\theta_{13}) \sim 0.01$, $\delta(\Delta m_{31}^2) < 1 \times 10^{-4}$ in the first stage and to determine the CP parameter in the second stage for values $|\delta_{CP}| > 10°$ with upgraded accelerator power (~1 → 4 MW). NOvA plans to have a second detector at a different off-axis angle to the beam line (i.e., different energy) in the second stage. Both groups have similar capability to determine the CP-violation parameter but NOvA has the additional potential to determine the sign of Δm_{31}^2 because of its long baseline. The correction due to matter effect is $2\beta \propto \Delta m_{31}^2$ [see Eq. (2.132b), (2.133)], which is negligible for T2K with $L = 295$ km but amounts to ~25% for NOvA with $L = 735$ km.

2.8.4

Future Prospects

Ongoing or scheduled experiments have limited capabilities in determining δ_{CP} and the sign of Δm^2_{31}. To increase the detection capability in case of unsuccessful current efforts and at the same time to plan precision experiments other than to determine the oscillation parameters, researches on possible high-quality, intense neutrino beams are under way.

Three Possible Facilities: SB, BB, and NF Three possibilities are discussed, the super-beam (SB), beta-beam (BB), and the neutrino factory (NF). Conventional way of making neutrino beams is (i) to produce π/K beam by bombarding primary proton on target, (ii) to focus the secondary beam using horns, and (iii) to obtain neutrinos through decays $\pi^{\pm}(K^{\pm}) \to \mu^{\pm} + \overset{(-)}{\nu}_{\mu}$. SB [274, 275, 276, 277, 278] increases the intensity of the neutrino beam by upgrading the accelerator proton beam power from $\lesssim$1 MW (=10^6 W) to ~4 MW. Upgrades of T2K, NOvA, and other approved experiments rely on this type of neutrino beam.

For instance, T2HK is an improved version of T2K with the SuperK (50 kt) upgraded to HyperK (1 Mt) placed at the present site at a distance of 295 km. T2KK is an alternative version of T2HK and divides the HyperK detector into half and places one of them in Korea at ~1200 km, improving the $\delta \sin^2 2\theta_{13}$ to 10^{-2}–10^{-3}. Upgrade of NOvA is also envisaged by placing the second detector at the second beam with a different production angle, producing a neutrino beam with a different energy.

Technically, SB is the most practical method because it is a scaled-up version of the established technique. Possible limitations are irreducible ν_e contamination due to Kl_3 and μ decays in the ν_μ beam which could overshadow the oscillation of $\nu_\mu \to \nu_e$. Ultimate limitation to this method is considered to be the understanding of the energy spectrum of pion production. Reaching the overall systematic error to the level of 5% or less is the key to successful experiments.

The beta beam [279, 280, 281] uses ν_e or $\overline{\nu}_e$ produced by beta-decay of stored radioactive beams. Typically, ^{6}He or ^{8}Li to produce ν_e and ^{18}Ne or ^{8}B to produce $\overline{\nu}_e$ are used. As the electron neutrino is produced via low-energy beta decay, there are no other background reactions in the rest frame of the parent nuclei. When the parent nuclei are accelerated, they provide a high-quality monochromatic ν_e beam, and an essentially background-free, pure $\nu_e \to \nu_\mu$ channel, an ideal setting for investigating the appearance of the wrong-sign muons.[29]

The energy of the neutrino beam is governed by $Q\gamma$, where Q is the neutrino energy in the rest frame of the parent nuclei and γ is the acceleration Lorentz factor. If one uses available facilities to accelerate the unstable nuclei $(\gamma_{\text{He}}, \gamma_{\text{Ne}}) = (150, 250)$ with CERN-SPS (Super-Proton-Synchrotron), $(350, 580)$ with refurbished SPS or Tevatron and (~2500, 4000) with LHC can be obtained. The γ's that could be

---

29) Pure $\nu_e \to \nu_\mu$ is often referred to as the *golden channel*, $\nu_e \to \nu_\tau$ as the *silver channel*, and $\nu_\mu \to \nu_e$ as the *bronze channel*.

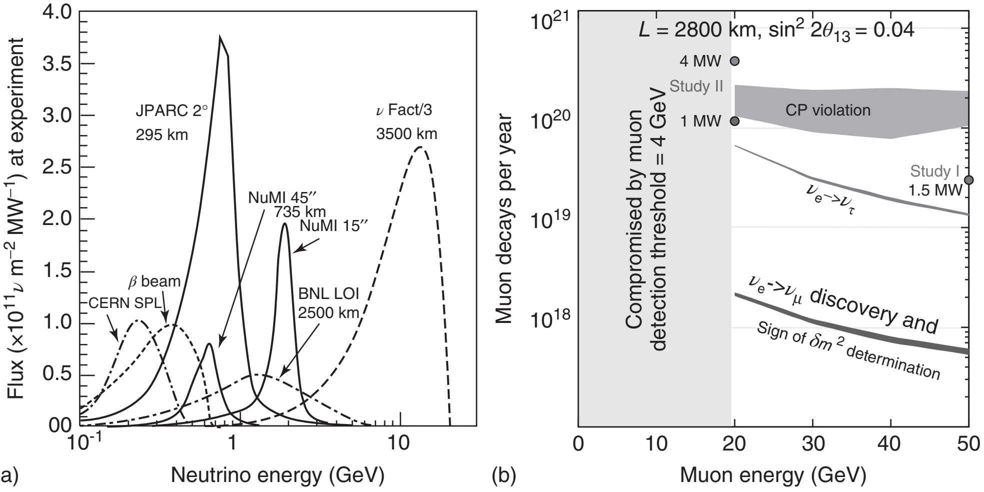


Figure 2.42 (a) Neutrino fluxes for variety of beam lines. Super beams, beta beams (ν_e), and neutrino factories are shown. The fluxes are normalized to 1-MW proton power, except for the beta beam, which uses considerably less proton power but is limited by different factors. Note that the energy of the NF beam is scaled down by a factor 3. (Reproduced with permission of [283].) (b) Muon decay rates of $10^{20} - 10^{21}$ per year can be expected for the neutrino factory. The drawing shows them in a straight section per 10^7 s (effectively one year) versus muon energy, with fluxes required for different physics searches assuming a 50-kt detector at $L = 2800$ km and $\sin^2 2\theta_{13} = 0.04$. (Reproduced with permission of [284].)

achieved for Li/B are $\gamma_{\text{Li/B}} = 8/9\gamma_{\text{He/Ne}}$. Therefore, the neutrino energy is in the range 0.75/1.5 GeV unless LHC is used. Simultaneous operation of the beta beam and the second-generation SB is also possible [282]. The disadvantage of the beta beam is the lack of the silver channel, the $\nu_e \to \nu_\tau$ transitions for the low-γ option. Figure 2.42a compares the beta beam with variety of super beams.

The NF [285, 286, 284, 276] derives an intense high-energy (20 – 45 GeV) neutrino beam from the decay of stored muons ($10^{20} - 10^{21}$ decay muons per year) and has access to all the channels of the neutrino-flavor transition including the golden channel. Its energy spectrum is also plotted in Figure 2.42a. As the energy is high, it requires the use of a magnetized detector possibly of the MINOS- or NOvA-type[30] calorimeter to reject beam-induced background reactions. Access to the silver channel ($\nu_e \to \nu_\tau$) can be facilitated by using either an emulsion-based detector (ECC) or a magnetized liquid-argon time-projection chamber [287]. Some physics examples envisaged for the NF experiment with the 50-ton detector are depicted in Figure 2.42b. Studies [288] so far have shown that the NF gives the best performance over virtually all of the parameter spaces, reaching down to a level of $\delta \sin^2 2\theta \simeq 10^{-5}$, which is to be compared with $\sim 10^{-2}$ envisaged by the super beam.

---

[30] If one opts for a large mass, a MINOS-type (Fe + scintillator/track detector) is preferred. If one opts for fine resolution, active target cum detector made of scintillators/track detector, that is, NOvA type, is preferred. A liquid argon detector is a possible alternative.

Besides, the NF also enables us to observe a variety of neutrino interactions which can be explored only with the highly intense neutrino beam. However, it requires acceleration of a muon beam which needs development of many novel accelerator techniques like durable targets, phase rotation, ionization cooling, and so on.[31] Its timescale and cost remain a major issue. Note that these developments are also required for the construction of a possible future muon collider beyond the LHC era.

Optimization of Future Experiments For the long-baseline accelerator experiments, the solar term containing Δm_{21}^2 is small, and an optimum condition is to set two detectors at a distance to make the atmospheric oscillation probability maximum, that is

$$\frac{\Delta m_{31}^2 L}{4E} = \frac{\pi}{2} \quad \rightarrow \quad E\,(\text{GeV}) = \frac{L}{515\,\text{km}} \left(\frac{\Delta m_{31}^2}{2.41 \times 10^{-3}\ \text{eV}^2} \right) \tag{2.135}$$

Most planned beams, except the neutrino factory which peaks at $\sim 30 - 40$ GeV, are in the neighborhood of $E_\nu \sim 1$ GeV (see Figure 2.42a).

Figure 2.43a shows the transition probability of $\nu_\mu \rightarrow \nu_e$ as a function of energy for $L = 2540$ km. Transition rates at other settings should simply scale with E/L, if there is no matter effect.[32] However, it becomes sizable for the long baseline. If one takes it into account, the functional shape changes considerably. Figure 2.43b shows the predicted ratio of $P(\bar{\nu}_e \rightarrow \bar{\nu}_\mu)/P(\nu_e \rightarrow \nu_\mu)$ as a function of the baseline. The CP effect nears its maximum around at $L \simeq 2000 - 4000$ km, which is an optimum setting for measuring the CP-violation effect, but the difference due to matter effect is larger for larger distances, which is a consideration if one wants to determine the mass hierarchy. Sign flip due to the mass hierarchy, that is, $X_{31} \rightarrow -X_{31}$ is equivalent to $\delta \rightarrow -\delta$ as one can see from Eq. (2.130), but the CP reversal changes also $\beta \rightarrow -\beta$ and can be distinguished from the sign flip of X_{31} alone. One needs a long baseline for CP as well as for mass hierarchy resolution, which requires high intensity. The NF is considered capable of delivering $10^{20} - 10^{21}$ decay muons per year (Figure 2.42b).

2.9
Double Beta Decay

The seesaw mechanism is the most plausible explanation for the small mass of the neutrino. A heavy Majorana neutrino is an essential ingredient not only of the seesaw mechanism but also of leptogenesis (see Section 8.6.3). If the neutrino is a

---

31) The phase rotation is an accelerator technique to convert a beam that has small Δt, wide ΔE (typical of the secondary particles) to small ΔE, wide Δt by using the phase space ($\Delta E - \Delta t$) volume invariance. The ionization cooling reduces transverse beam spread by repeated application of ionization loss through matter and acceleration. The conventional cooling technique (stochastic or electron) is too slow to be applied to fast decaying muons.

32) The graph is shown for a canonical setting for the NF. E_ν is slightly less than E_μ. Therefore, the scaling factor should be smaller than 20, but we ignore the difference for qualitative argument.

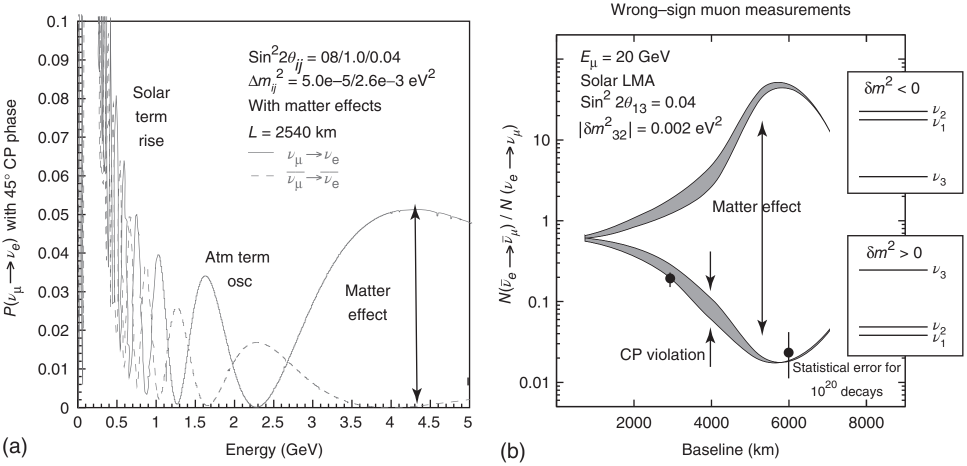


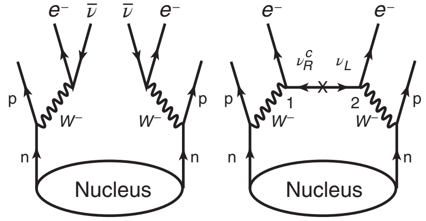
Figure 2.43 (a) Probability of ν_μ oscillating into ν_e after 2540 km. This plot assumes a CP-violation phase of +45°. The upper and lower curves are for neutrinos and anti-neutrinos, respectively. Violent oscillation below $E_\nu \lesssim 1$ GeV is due to the solar term (i.e., X_{21}). The first and second (from the right), and so on, conspicuous peaks of the atmospheric term (i.e., X_{31}) are seen. For the assumed distance of 2540 km, the matter effect is large and leads to almost complete reversal of nodes and antinodes for neutrinos and anti-neutrinos. The probability for neutrinos with $\Delta m^2_{32} < 0$ is similar to (but not exactly the same as) that of anti-neutrinos. (Reproduced with permission of [289].) (b) $(\nu_e \to \nu_\mu)$ rates when positive and negative muons are stored in a 20-GeV neutrino factory, shown as a function of baseline. A muon measurement threshold of 4 GeV is assumed. The lower and upper bands correspond, respectively, to negative and positive Δm^2_{32}. The width of the bands shows how the predictions vary as the CP-violating phase δ is varied from $-\pi/2 \to \pi/2$. Arrows at 4000 and ~5200 km are the optimum distances for measuring CP violation and for discriminating the mass hierarchy. Statistical errors correspond to a high-performance neutrino factory yielding a data sample of 10^{21} decays with a 50-kt detector. (Reproduced with permission of [284, 290, 291].)

Majorana particle, it means the lepton number is violated. At present, the double beta decay is the only reaction that can be used to test the Majorana neutrino [292, 293]. It is a process in which a nucleus emits two electrons and converts two of its constituent neutrons to protons. Two modes exist, 2ν and 0ν.

$$2\nu \text{ mode}: \qquad (Z) \to (Z+2) + 2e^- + 2\overline{\nu}_e \tag{2.136a}$$

$$0\nu \text{ mode}: \qquad (Z) \to (Z+2) + 2e^- \tag{2.136b}$$

In the 2ν mode, an ordinary beta decay occurs twice inside the nucleus (Figure 2.44a) and is expected as a higher order process in the Standard Model. In the 0ν mode, the lepton number changes by 2 before and after the decay and the process occurs only when the conservation law is violated. Take a look at Figure 2.44b. This is the process in which a neutrino emitted at the vertex 1 and another emitted at the vertex 2 annihilates together. In the conventional interpretation, one neutrino is emitted at one end and the other is absorbed by the



(a) (b)

Figure 2.44 Diagrams for double beta decay. (a) 2ν decay. (b) $0-\nu$ decay.

electron. So what is special about it? In the Standard Model, if a neutrino is emitted together with an electron, it is the anti-neutrino and right-handed because of the V−A nature of the weak interaction. The same V−A interaction constrains that the neutrino can be converted to an electron only when it is in a particle state and is left-handed. Therefore, the emitted anti-neutrino can be converted to an electron only if (i) the lepton number is violated and (ii) either the neutrino has a finite mass or a nonstandard V+A interaction exists. This is because the massive right-handed neutrino can have a left-handed component (if massive, there exists a Lorentz transformation to flip the handedness) and can be absorbed via a V−A interaction or its right handed-component is absorbed directly via a V + A interaction.

2.9.1
The Effective Majorana Mass

The probability of having the opposite helicity component is proportional to $(m_\nu/E)^2$ (see Section 4.3.5 of [1]); therefore, if the 0ν double beta decay is detected, one can also determine the neutrino mass. The decay rate is proportional to the square of the nuclear matrix element $|M^{0\nu}|^2$, the phase space volume ($G^{0\nu}$), and the square of the effective mass $<m_{\beta\beta}>^2$.[33]

$$(T_{1/2}^{0\nu})^{-1} = G^{0\nu}|M^{0\nu}|^2 <m_{\beta\beta}>^2 \tag{2.137}$$

where $T_{1/2} = \tau/\ln 2$ is half-lifetime when one-half of the original material has decayed. Theoretical uncertainty of the nuclear matrix is up to a factor of 10, which causes a spread of the measured values of the half-lifetime. The effective mass is a superposition of the mass eigenvalues and is given by

$$\begin{aligned}<m_{\beta\beta}> &= \sum_j U_{ej}^2 m_j \equiv \sum_j m_j \eta_j |U_{ej}|^2 \\ &= m_1 c_{12}^2 c_{13}^2 + m_2 s_{12}^2 c_{13}^2 e^{i(\alpha_2-\alpha_1)} + m_3 s_{13}\ 2e^{-i(\alpha_1+2\delta)} \quad {}^{34)}\end{aligned} \tag{2.138}$$

where U_{ej}'s are the mixing matrix elements and η_j is the CP phase factor which changes its sign depending on the CP eigenstate. Notice that, if CP is not violated (i.e., $\delta = 0$, $e^{i\alpha_1}$, $e^{i\alpha_2} = \pm 1$), then η_j becomes ± 1. Depending on the relative magnitude of the three masses, the effective mass $<m_{\beta\beta}>$ takes a different value. Of special interest are the following three extreme cases where we have set $\Delta m^2_\odot \equiv \Delta m^2_{21}$, $\Delta m^2_A \equiv \Delta m^2_{31} \simeq \Delta m^2_{32}$:

---

[33] We use $<m_{\beta\beta}>$ to denote the neutrino mass obtained in the double beta decay.

[34] Notice that the factor U_{ej}^2 appears instead of $|U_{ej}|^2$. This is because the mass Lagrangian of the Majorana particle is of the form $\sim\psi_L^T C\psi_L$ [see Eq. (2.6)] while that of Dirac particle is $\sim\overline{\psi}\psi$.

(1) Quasi-degeneracy (QD)

$$|m_1| \simeq |m_2| \simeq |m_3| = m_0 \tag{2.139}$$

(2) Inverted hierarchy (IH): $m_2 > m_1 > m_3$

$$m_2 = \sqrt{m_3^2 + \Delta m^2_\odot - \Delta m^2_A}, \qquad m_1 = \sqrt{m_3^2 - \Delta m^2_A} \tag{2.140}$$

(2.141)

In the extreme case,

$$|m_2| \simeq |m_1| \simeq \sqrt{|\Delta m^2_A|} \gg m_3 \tag{2.142}$$

(3) Normal hierarchy (NH): $m_3 > m_2 > m_1$:

$$m_2 = \sqrt{m_1^2 + m^2_\odot}, \qquad m_3 = \sqrt{m_1^2 + \Delta m^2_A} \tag{2.143}$$

In the extreme case,

$$|m_3| \simeq \sqrt{|\Delta m^2_A|} \gg m_2 \simeq \sqrt{|\Delta m^2_\odot|} \gg m_1 \tag{2.144}$$

As $\Delta m^2_\odot$ and $|\Delta m^2_A|$ are known from the oscillation experiments, one can express the two heavier masses in terms of the lightest mass, that is, m_1 for the NH and m_3 for the IH. Since θ_{12} and θ_{13} are also known, the value of the effective mass $<m_{\beta\beta}>$ can be determined as a function of the lightest mass if the phases are fixed. Uncertainty of the phases expands a line to a band with CP phase $\eta_j = \pm 1$ as the extreme case, giving maximum or minimum of the tolerance band.

The three branches are depicted in Figure 2.45 [294].

For the special case in which all the three terms are of comparable size, the effective mass can be very small, which is indicated by the upward-going branch at the upper right corner in Figure 2.45. One sees that minimum value of the heaviest neutrino for IH is within the band $<m_{\beta\beta}>= 0.01 - 0.05$ eV and larger for quasi-degeneracy (QD). With currently available technology, this value can be reached in the near future.

2.9.2
Current Status

Conceptually, it is easy to distinguish the 0ν from the 2ν mode. Sum of the two electron energies is fixed for the 0ν mode and its spectrum should be monochromatic, whereas it is continuous in the 2ν mode (see Figure 2.46A).

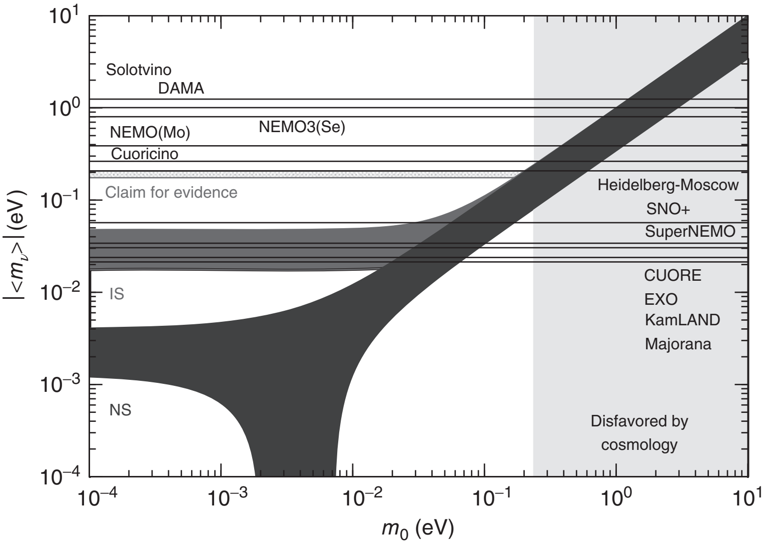


Figure 2.45 Allowed range of values for $|<m_{\beta\beta}>|$ as a function of the lowest mass m_0 using the three standard neutrinos for the cases of normal spectrum (NS, $m_0 = m_1$) and inverted spectrum (IS, $m_0 = m_3$) of neutrino masses. Also shown are the current experimental limits and the expected future results. [295] Note that in the IH, there is a lower bound, which means that in such a scenario the $0\nu\beta\beta$ decay should definitely be observed if the experiments reach the required level. (Reproduced with permission of [296, 294].)

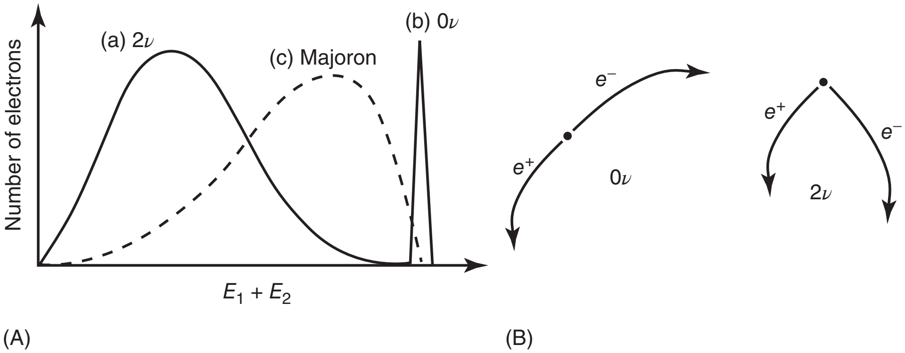


Figure 2.46 Characteristics of double beta decays. (A) Energy spectrum. (B) track topology.

Besides, if tracks of the emitted electrons are recorded, they are back to back in the former while they can make any angle in the latter (Figure 2.46B). Nuclei that undergo the double beta decay generally have lifetimes $\gtrsim 10^{18-21}$ years or more. Therefore, the decay is a rare process and experiments to measure the lifetime suffer from environmental backgrounds (cosmic rays and/or isotopes that exist in the detector material). The 2ν modes have already been observed [8]. Traditionally, the lifetime is expressed in $T_{1/2}$ (half-lifetime). Table 2.4 is a partial

Table 2.4 Measured half-lifetime from 2ν double beta decay.

Nuclide	$T^{2\nu}_{1/2}$ yr [297, 293]	$T^{2\nu}_{1/2}$(theory) [298]
^{76}Ge	$1.5 \pm 0.1 \times 10^{21}$	$(0.07 - 60) \times 10^{21}$
^{100}Mo	$7.1 \pm 0.4 \times 10^{18}$	$(0.1 - 2 \times 10^{4}) \times 10^{18}$
^{116}Cd	$3.0 \pm 0.2 \times 10^{19}$	$(0.3 - 200) \times 10^{19}$
^{150}Nd	$7.8 \pm 0.7 \times 10^{18}$	$(0.06 - 400) \times 10^{18}$

Table 2.5 Upper limits of half-year and m_ν from 0ν double beta decay [7, 293].

Nuclide	$T^{0\nu}_{1/2}$ yr 90% CL	$m_{\beta\beta}$ (eV) 90% CL	
76Ge	$2.23^{+0.44}_{-0.31} \times 10^{25}$	0.32 ± 0.03	[299]
^{100}Mo	5.8×10^{23}	$<0.45 - 0.93$	[302]
^{130}Te	1.8×10^{24}	$<0.19 - 0.68$	[303]
^{136}Xe	1.6×10^{25}	$<0.14 - 0.38$	[300, 301]

list of observed, averaged, and recommended values. As stated before, there are considerable ambiguities in estimates of the nuclear matrix elements.

However, no 0ν mode has been detected so far.[35] We list some of recent results together with the upper limits on the mass value in Table 2.5. The range of values reflects the ambiguities of the nuclear matrix elements.

Given the large uncertainties in the value of the nuclear matrix elements, one might worry how reliable are the values of the effective mass. In recent years, the calculation of the nuclear matrix elements has been much improved, and the discrepancy much reduced. For the most thoroughly investigated germanium, it is argued that the differences are less than a factor 10, which means that the effective mass ambiguity is within a factor of 3 [304]. However, it is likely that it will remain as the largest ambiguity factor.

One naturally expects to discover at least one of the neutrinos soon. The maximum necessary sensitivity with present prediction is ~0.01 eV. Many proposals are being made to realize a ~1-t detector with sensitivity $<m_{\beta\beta}>$ ~0.03 – 0.1 eV [293].

2.9.3
To Design an Experiment

Since the double beta decay is a rare process, it is very hard to obtain positive signals and, if they are there, one must convince other investigators that they are real and not fake events. We discuss briefly some of the experimental difficulties associated with such rare events.[36] A good experiment has a good signal-to-noise ratio with enough statistics. Let us take a look at how to maximize the number of events and how the background can be reduced.

---

[35] At the time of writing, one group has claimed the detection of 0ν mode in ^{76}Ge and derived a finite value of $m_\nu = 0.32 \pm 0.03$ eV [299]. However, it is not confirmed by subsequent experiments [300, 301].

[36] Arguments in this section owe much to [305, 306].

Choice of isotopes: First, we have to choose an isotope with lifetime as short as possible. Since the lifetime is inversely proportional to the phase space volume [see Eq. (2.137)], an isotope with a high Q-value is desired. Considering compatibility with detector techniques, the following isotopes are usually used.

^{48}Ca (4.27), ^{76}Ge (2.04), ^{82}Se (3.00), ^{96}Zr (3.35), ^{100}Mo (3.03), ^{116}Cd (2.81), ^{130}Te (2.53), ^{136}Xe (2.46), and ^{150}Nd (3.37).

Their Q-values are given in parentheses. They are referred to as the *magnificent nine*. In addition, isotopic abundance and ease of enrichment are also required. For instance, ^{48}Ca has by far the largest Q-value, but the natural abundance of the isotope is only 0.18% and is technically hard to enrich. Other isotopes have typically ~10% natural abundance.

Figure of Merit: In an experiment where a signal has to be sought on a large, flat background, its reduction is one of the prime considerations. The number of total signal events from the double beta decay is given by

$$N_{\text{event}} = \frac{\ln 2}{T_{1/2}} \times \epsilon \cdot N_{\text{decay}} \cdot t_{\text{meas}} \tag{2.145}$$

$$N_{\text{decay}} = \frac{\eta N_A M}{A}$$

Here the variables denote $T_{1/2}$: half life, t_{meas}: measured time, ϵ: detector efficiency, η: the ratio between the total mass and the detector (source) mass, N_A: Avogadro number, M: detector fiducial mass, and A: the atomic number of the source. The background is given by

$$N_{\text{bkg}} = \text{bkg} \cdot \Delta E \cdot M \cdot t_{\text{meas}} \tag{2.146}$$

where "bkg" means number of background/energy bin/s, and ΔE denotes the energy resolution. Then, the Figure of Merit (FOM) can be expressed as signal to noise ratio (S/N)

$$\text{FOM} = \text{S/N} = \frac{N_{\text{event}}}{\sqrt{N_{\text{bkg}}}} \propto \epsilon \cdot \frac{\eta}{A}\sqrt{\frac{M \cdot t_{\text{meas}}}{\text{bkg} \cdot \Delta E}} \tag{2.147}$$

As is obvious from the above expression, one needs

- good energy resolution
- low background, and
- large mass and long measurement time

to obtain a good S/N. In addition to the flat background, decay events from $2\nu\beta\beta$ in the upper end of the spectrum has to be discriminated. The ratio $R_{0\nu/2\nu}$ of counts due to $0\nu\beta\beta$ decay over those due to $2\nu\beta\beta$ in a narrow window around the Q-value (of the order of the detector energy resolution) is given by the following expression [305]:

$$R_{0\nu/2\nu} = \frac{m_e}{7Q\delta^6}\frac{T_{1/2}^{2\nu\beta\beta}}{T_{1/2}^{0\nu\beta\beta}}, \qquad \delta = \frac{\Delta E}{Q} \tag{2.148}$$

Notice the strong dependence of the ratio on the energy resolution. A slow $2\nu\beta\beta$ decay rate (like ^{136}Xe, for which $T_{1/2}^{2\nu\beta\beta} = 2.2 \times 10^{21}$ years) helps more than a fast decay rate (like ^{100}Mo, for which $T_{1/2}^{2\nu\beta\beta} = 7.1 \times 10^{18}$ years). For the latter, an excellent energy resolution (<1%) is mandatory.

Smaller $<m_{\beta\beta}>$: If one wants to put more emphasis on the value of $<m_{\beta\beta}>$ rather than on good measurement of the lifetime, one has to combine Eqs. (2.147) and (2.137). For a fixed FOM one obtains

$$<m_{\beta\beta}>= \frac{1}{(G^{0\nu})^{1/2}|M^{0\nu}|(T_{1/2}^{0\nu})^{1/2}} \propto \frac{1}{(G^{0\nu})^{1/2}|M^{0\nu}|}\left[\frac{\text{bkg}\cdot\Delta E}{M\cdot t_{\text{meas}}}\right]^{1/4} \tag{2.149}$$

The expression shows a very weak dependence of the sensitivity on the set-up parameters. For instance, if an experiment has established $<m_{\beta\beta}>= 0.2$ eV with 100 kg of mass, it has to be increased by a factor $4^4 = 256$ to reach 0.05 eV. One has to think hard to increase the sensitivity.

So far, we have considered only background limited experiments. The flat background is unavoidable if one measures only the energy sum of the two emitted particles as is the case for a detector consisting of Ge calorimeter only, which has been a norm in the early stage of double beta decay measurements. If one can combine energy and track information together, and further if the daughter nuclei can be identified, it may be possible to realize a zero-background investigation. In this case, the statistics is not a Gaussian. One has to use Poisson statistics. An observation of 0 counts means the background number $N_{\text{bkg}} = 3$ for 95% confidence level. In this case, FOM is given by

$$\text{FOM} = \frac{\text{N}_{\text{event}}}{N_{\text{bkg}}} \overset{(2.145)}{=\!=} \frac{\ln 2}{T_{1/2}^{0\nu}}\frac{N_A}{A}\epsilon\cdot\eta\frac{M\cdot t_{\text{meas}}}{N_{\text{bkg}}} \tag{2.150}$$

and Eq. (2.149) is modified to

$$<m_{\beta\beta}> \propto \frac{1}{(G^{0\nu})^{1/2}|M^{0\nu}|}\frac{1}{\sqrt{M\cdot t_{\text{meas}}}} \tag{2.151}$$

Equation (2.151) shows that for a detector that has realized zero background, mass scalability can be applied to reach the desired value.

2.9.4
Experimental Apparatus

From the apparatus point of view, there are two extreme approaches: one uses calorimetry only where the source plays a dual role of the detector, and the other uses a tracking device at the expense of energy resolution. The most traditional and still a viable approach today is to use stacked Ge diodes to form a high-resolution (up to 0.1%) calorimeter. Examples of advanced apparatus are GERDA [307, 308] and MAJORANA [309, 310] which aim to obtain ~2×10^{26} years in the first stage and up to 6×10^{27} years in the second stage. The group that claimed a positive evidence [299] also used Ge. CANDLE [311], which takes advantage of the high Q-value of Ca using CaF$_2$ as scintillating detectors, also belongs to this category. Another method in this category is to use a bolometer, which is a calorimeter to measure the total energy released in the form of thermal phonons. An example is CUORE [312, 313], which uses TeO_2 bolometers with 0.25% resolution. Examples of tracking detectors are SuperNEMO [314] and NEXT [315]. The former is an extension of the successful NEMO3 [316, 317], which uses an external source and magnet to obtain superb tracking capability to reach virtually zero background but with low energy resolution. The latter uses a gaseous-xenon time projection chamber (TPC) to obtain track reconstruction capability as well as high energy resolution of ~1%.

Finally, we discuss two detectors that have scalability to large mass with potential to realize the zero background environment and hence the capability to reach the 10-meV scale.

One is EXO (enriched Xenon observatory), which by using liquid-xenon TPC [318] gains a large source mass. The TPC's major advantage is its capability to reconstruct three-dimensional topology as well as scintillation/ionization energy information realizing virtually zero background. By using the liquid, it loses tracking information but still retains its advantage of 3-D reconstruction capability. Eventually, it is hoped to develop to a multi-ton detector with barium tagging (i.e., tag the daughter nuclei of ^{136}Xe), reaching sensitivity of Majorana mass below 10 meV. EXO200, which is its small version, currently leads the field.

The other is KamLAND-Zen [300] with large ultra radio-pure LSs which were initially used for neutrino oscillation measurements. They were converted to $\beta\beta$ detector by adding an isotope Xe-loaded double beta decay source inside the LS covered by a thick active shield which was the former KamLAND detector. It is a low-resolution calorimeter-type detector. Its merit is its scalability, that is, the ease with which to reach quickly a large amount of isotope if background suppression is successfully achieved.

The above list comprises just a few among the many $\beta\beta$ detectors. Some of the claimed mass reaches are indicated in Figure 2.45. The reader may refer to recent review talks for more details [306, 296, 319].

EXO200 EXO200 is a TPC [320]) using enriched (purity 80.6% ^{136}Xe, 19.4% ^{134}Xe) liquid Xe (xenon) to serve as both the source of the nuclear decays and the detection material (see Figure 2.47). Xe has a Q-value of 2457.83 ± 0.37 keV [301]. Energy deposition in the liquid Xe produces both ionization and scintillation signals (see Figure 10.22). The TPC is capable of three-dimensional topological and temporal reconstruction of individual energy deposition. Together with the capability for obtaining dual energy information, it can also discriminate $\beta\beta$ decays from residual backgrounds dominated by γs. Calibration data were collected with three γ sources [^{228}Th (2615 keV)] at three positions inside the TPC.

The ability of the TPC to identify multi-site events (MS: mostly β decays) and single-site events (SS: mostly $\beta\beta$ decays) are used to separate multiple-site γ interactions.

Figure 2.48 shows the energy spectra of SS events as measured by the ionization and scintillations channels. ^{228}Th source was deployed to calibrate the energy. The Xe $2\nu\beta\beta$ spectrum is shown as shaded area. Contributions of backgrounds

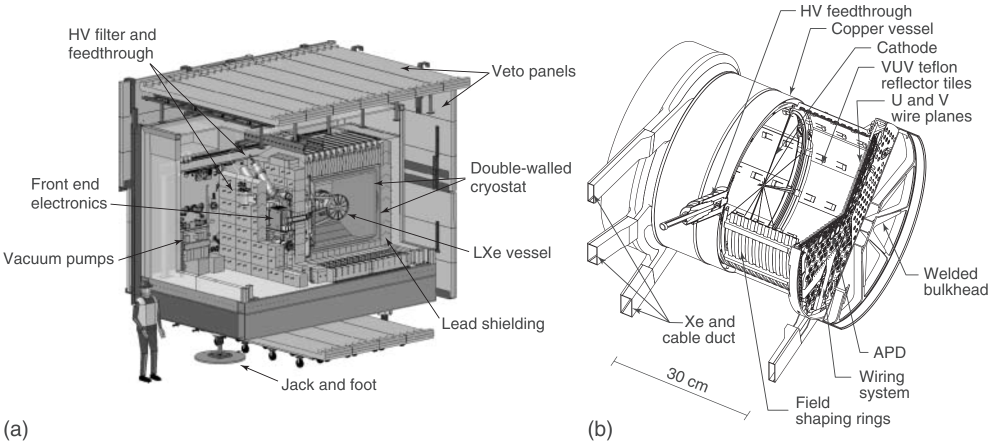

HV filter and feedthrough
Veto panels
Front end electronics
Double-walled cryostat
Vacuum pumps
LXe vessel
Lead shielding
Jack and foot
(a)
HV feedthrough
Copper vessel
Cathode
VUV teflon reflector tiles
U and V wire planes
Welded bulkhead
Xe and cable duct
30 cm
APD
Wiring system
Field shaping rings
(b)

Figure 2.47 (a) Cutaway view of the EXO-200 setup, with the primary subassemblies identified. The TPC is mounted in the center of a low-background cryostat. At least 50 cm of HFE-7000 fluid (HFE-7000 is a heat transfer fluid pumpable to −120° [321].) and additional 25 cm of lead shield the TPC from external radioactivity. The clean room housing the TPC is surrounded on four sides by 50-mm-thick plastic scintillator array veto panels. The whole apparatus is located at the depth of 1585 m W.E. at WIPP in New Mexico, USA. (b) Cutaway view of the EXO-200 TPC with the main components identified. It has a geometry of a cylinder of 40 cm diameter and 44 cm length, with a cathode grid dividing the cylinder into two identical regions. The cathode bias was set at −8.0 kV, providing a drift field of 376 V cm^{-1}. Ionized electrons drift toward the end cap to which are attached wire grids and an array of large-area avalanche photodiodes (LAAPD), which allow simultaneous readout of ionization and scintillation in liquid Xe. The wire grids provide two-dimensional localization and energy readout of each charge deposition. The third (longitudinal) coordinate is obtained from the time interval between the scintillation signal in LAAPDs and the collection of charge at grids. (Reproduced with permission of [318].)

are also shown in Figure 2.48a. The energy resolution σ/E was 1.67%. Only 1 (5) events were observed within $\pm 1\sigma(\pm 2\sigma)$ regions around the Q value (denoted as $Q_{\beta\beta}$) (see Figure 2.48b), hence no evidence for $0\nu\beta\beta$ decay was found. The fiducial volume used in the analysis contains 79.4 kg of ^{136}Xe (3.52×10^{26}) atoms giving the background rate of (1.5 ± 0.1/keV year t). A fit produces a lower limit of $T_{1/2}^{0\nu\beta\beta}$ as well as a value of $T_{1/2}^{2\nu\beta\beta}$.

$$\begin{aligned} T_{1/2}^{0\nu\beta\beta} &> 1.6 \times 10^{25}\ \text{years} \quad 90\,\%\ \text{CL} \\ T_{1/2}^{2\nu\beta\beta} &= (2.23 \pm 0.017\,\text{stat} \pm 0.22\,\text{syst}) \times 10^{21}\ \text{years} \end{aligned} \tag{2.152}$$

This value corresponds to the upper limit of the averaged Majorana mass

$$< m_{\beta\beta} > \, < 140 - 380\,\text{meV} \tag{2.153}$$

depending on the value of nuclear matrix elements.

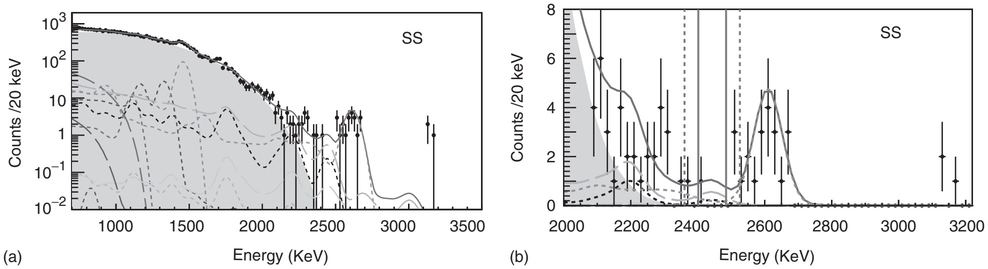


Figure 2.48 (a) Energy spectrum of single-site (SS) events obtained by EXO-200. The best-fit line (solid blue) is shown. The background components are $2\nu\beta\beta$ (gray region), ^{40}K (dotted orange), ^{60}Co (dotted dark blue), ^{222}Rn in the cryostat–lead airgap (long-dashed green), ^{238}U in the TPC vessel (dotted black), ^{232}Th in the TPC vessel (dotted magenta), ^{214}Bi on the cathode (long-dashed cyan), ^{222}Rn outside of the field cage (dotted dark cyan), ^{222}Rn in active xenon (long-dashed brown), ^{135}Xe (long-dashed blue), and ^{54}Mn (dotted brown). (b) Energy spectra in the ^{136}Xe $Q_{\beta\beta}$ region for SS events. The $1(2)\sigma$ regions around $Q_{\beta\beta}$ are shown by solid (dashed) vertical lines. The peak at 2615 keV is due to ^{228}Th. The fit results have the same meaning as in (a). The $0\nu\beta\beta$ probability density function (PDF) from the fit is not visible. (Reproduced with permission of [301].) (Please find a color version of this figure on the color plates.)

KamLAND-Zen The KamLAND-Zen (zero neutrino) experiment consists of 13 tons of xenon-loaded liquid scintillator (Xe-LS) contained in a 3.08-m-diameter transparent nylon-based inner balloon (IB), suspended at the center of the KamLAND detector by film straps. The IB is surrounded by 1 kt of LS contained in a 13-m-diameter outer balloon. To detect scintillation light, 1325 (17-in.) and 554 (20-in.) PMTs are mounted on the stainless-steel containment tank providing 34% coverage. The stainless-steel containment tank is surrounded by a 3.2-kt water-Cherenkov detector for cosmic-ray muon identification. Details of the KamLAND-Zen detector are given in Ref. [300]. They achieved the result comparable to the result with EXO200. They observed a rather large peak at the place exactly where 0ν beta decay of Xe should appear (see Figure 2.49). But its shape was wrong, and the group interpreted it as backgrounds coming mainly from $^{110\text{m}}$Ag in the LS. This is a typical example where a new type of detector always faces a danger of suffering from unexpected backgrounds and discovery of a monochromatic peak alone does not warrant detection of the true signal. Although the KamLAND-Zen detector has advantage of easy scalability to reach the desired sensitivity, its success critically depends on its ability to suppress the backgrounds.

2.10
Supernova Neutrino

The explosion of supernova 1987A in the Magellanic Cloud was a historic event. It was the first supernova visible to human eyes since Kepler. Two groups, Kamiokande and IMB, also observed the neutrino bursts for the first time (Figure 2.50) [323, 324, 325, 326, 327, 328]. Together with observation of solar

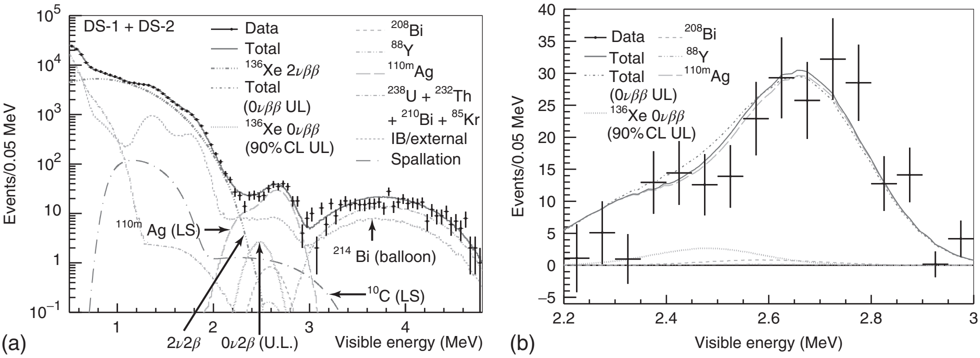


Figure 2.49 (a) Energy spectrum of selected candidate events by KamLAND-Zen together with the best-fit backgrounds and $2\nu\beta\beta$ decays, and the 90% CL upper limit for $0\nu\beta\beta$ decays. The fit range is $0.5 < E < 4.8$ MeV. (b) Closeup of (a) for $2.2 < E < 3.0$ MeV after subtracting known background contributions. (LS) means backgrounds from liquid scintillator and (balloon) from balloon. (Reproduced with permission of [322].) (Please find a color version of this figure on the color plates.)

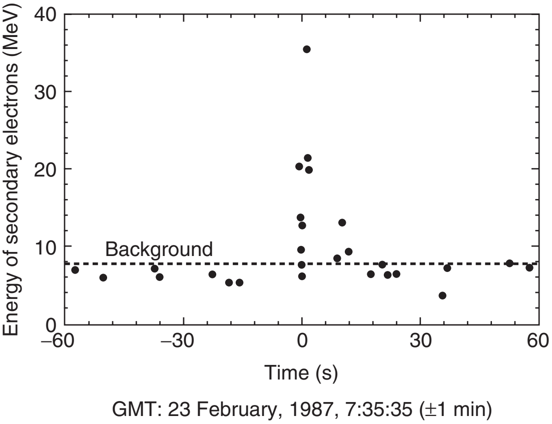


Figure 2.50 Observed supernova SN1987A events by Kamiokande. Eleven events are seen starting at t = 0 at the center. Flat low-energy events before and after $t = 0$ are backgrounds. (Reproduced with permission of [332, 333].)

neutrinos [329, 330], it heralded the dawn of neutrino astronomy. It was an epoch-making event in astrophysics because it confirmed basically the validity of the stellar evolution model which so far had only indirect evidences from light emission spectra. From the particle physics point of view, it is also an eye-opener because it has provided a new test bench otherwise impossible with terrestrial experiments. We will first learn the qualitative features of stellar evolution to understand the neutrino emission mechanism to further investigate the neutrino properties [131, 331].

2.10.1
Stellar Evolution

Most stars including the Sun finish their lives in the form of white dwarfs and fade away quietly. But large stars having mass more than eight times the solar mass (denoted as $M_\odot$), generally explode with spectacular fireworks, referred to as *supernovae*. They come in two kinds, type Ia and others (type II, Ib, Ic, etc.). The former explodes by deflagration[37] of carbons, oxygens, and so on and leaves nothing behind. It occurs when the star has a companion (i.e., when it is in a binary system). The latter explodes by core collapse/rebound and leaves behind a neutron star or a black hole. Two key concepts are worth noting in understanding the fate of the star. The first is the Virial theorem applied to particles in the gravitational potential which drives the stellar evolution to its death. The second is the degenerate pressure. As the star evolves, it synthesizes heavier elements progressively by burning lighter elements inside the hot and dense interior. Its final destiny is critically dictated by the timing of igniting a new fuel and of the star core being degenerate or not.

Protostar A star is born out of gas and dust that exists between the stars, known as the *interstellar medium* (*ISM*). Because of the universal attractive force of gravity, a core (the protostar) is formed out of a giant molecular cloud. It grows by absorbing surrounding materials (particles) that fall into the core, acquiring kinetic energy, and thermalize by collisions with other particles. The thermal motion produces pressure, which eventually counterbalances the gravity, and the whole protostar becomes quasi-stable. The pressure prevents collapse of the protostar and keeps its finite size in thermal equilibrium. However, according to the Virial theorem, half of the acquired energy is lost by emission of light from surface and, to supplement the lost energy, the protostar releases gravitational energy by contracting itself. The Virial theorem under gravity is given by the following formula, which applies to an ensemble of particles [see Eq. (A.9)].

$$K = U = -\frac{1}{2}\Omega = -E, \qquad E = U + \Omega = \frac{1}{2}\Omega < 0 \tag{2.154}$$

Here, K is the total kinetic energy, U the internal energy, Ω the gravitational potential energy, and E is the total energy. This is an important relation to understand stellar evolution. Equation (2.154) tells us that the star can produce energy by contraction, decreasing its potential energy and driving itself to a larger total negative energy state. It consumes half of the produced energy as radiation and stores the other half as the internal energy. The increase of the internal energy means the increase of temperature and pressure. As the temperature rises by losing energy, the star is a special system that has negative specific heat. To compensate

---

37) A special type of explosion akin to conventional fire burning. It propagates through thermal conductivity. It is a subsonic combustion to differentiate from detonation which is supersonic and propagates through shock compression.

the emitted energy, the star keeps contracting incessantly, thereby increasing the temperature, density, and pressure.

When the surface temperature exceeds 2000 K, the radiation enters the visible light range and begins to shine. This is the birth of a star. It keeps shining as long as energy is provided. A rough estimate of the time for how long it can keep shining can be obtained by assuming a constant density $\rho = \overline{\rho}$ and constant radiation. A sphere of mass M with uniform mass density distribution with radius R has the gravitational energy of $\Omega = -(3/5)M^2/R$, hence total mechanical energy $E = -(3/10)M^2/R$. Then, the energy of a star having the solar mass which will be radiated away during the collapse is given by

$$\Delta E = -(E_f - E_i) \simeq -E_f \simeq \frac{3}{10} G_N \frac{M_\odot^2}{R_\odot} \simeq 1.2 \times 10^{41}\ \text{J} \tag{2.155}$$

Assuming also that the luminosity has been roughly constant throughout its lifetime, the star could keep radiating at that rate for the duration of approximately

$$\tau_{\text{KH}} = \frac{\Delta E}{L_\odot} \sim 10^7\ \text{yr} \tag{2.156}$$

where $L_\odot$ is the Sun's luminosity. τ_{KH} is known as the Kelvin–Helmholtz timescale.

We show that the pressure P_c, density ρ_c, and temperature T_c at the stellar core are governed by global parameters, namely its total mass M and radius R. An order estimate of the state variables at the core can be obtained by approximating them with their average. A balance between the pressure and gravity can be obtained as

$$4\pi R^2 \overline{P} \simeq G_N \frac{M^2}{R^2} \qquad \rightarrow \qquad P_c \propto G_N \frac{M^2}{R^4} \tag{2.157a}$$

$$\rho_c \simeq \overline{\rho} = \frac{M}{4/3\pi R^3} \qquad \rightarrow \qquad \rho_c \propto \frac{M}{R^3} \tag{2.157b}$$

Eliminating R from the above two equalities, we get

$$\frac{P_c^3}{\rho_c^4} \propto G_N^3 M^2 \tag{2.158}$$

For an ideal gas, we have the equation of state

$$P_c = n k_B T_c = \frac{k_B}{\mu m_N} \rho_c T_c \tag{2.159}$$

where n is the average number density, k_B is the Boltzmann constant, μ is the average molecular weight, and m_N is the nucleon mass. Substituting Eq. (2.159) in Eq. (2.158), we have

$$\frac{T_c^3}{\rho_c} \propto G_N^3 M^2 \tag{2.160}$$

Degenerate Pressure Ordinary pressure is generated by thermal motion of particles and sensitive to the temperature. Degenerate pressure, on the other hand, is a quantum mechanical phenomenon and is almost independent of the temperature. Let us consider the electron gas. The electron is a fermion and, because of the Pauli's exclusion principle, no two particles can occupy the same state. In the configuration space, they cannot be in the same position. If they come close to each other, the Heisenberg uncertainty principle ($\Delta x \Delta p \gtrsim \hbar$) dictates that their momentum go up. The average momentum each particle has goes like $\overline{p} \sim 1/(\Delta V)^{1/3} \sim \rho^{1/3}$, and the average energy of the nonrelativistic particles $\varepsilon \propto \overline{p}^2/2m_e \sim \rho^{2/3}$. The average energy density is given by $u = n_e \varepsilon = (N_e/V)\varepsilon$. The pressure is given by

$$P_{nr} = \frac{dU}{dV} = \frac{d(uV)}{dV} = \frac{d(uM/\rho)}{d(M/\rho)} \simeq \rho^2 \frac{d\rho^{2/3}}{d\rho} \propto \rho^{5/3} \tag{2.161a}$$

Nonrelativistic particles will become relativistic eventually as the distance between particles becomes smaller. By the same logic, the pressure of extremely relativistic ($p \gg m_e$) particles ($\varepsilon \sim \overline{p}$) is given by

$$P_{er} \propto \rho^{4/3} \tag{2.161b}$$

If one inserts Eq. (2.161b) into Eq. (2.158), one sees that the total mass of the extremely relativistic and degenerate gas sphere is a constant. That is, there is an upper limit for the mass that the degenerate pressure can support. It is referred to as the *Chandrasekhar mass*. Its precise value depends on the chemical composition of the constituent particles but for degenerate electron gas, it is approximately given by [see Eq. (B.59)]

$$M_{\text{Ch}} = 1.46 \left(\frac{2}{\mu_e} \right)^2 M_\odot \tag{2.162}$$

where μ_e denotes the average number of nucleons per electron.

Evolutionary Paths: As a start, for considering stellar evolution, we temporarily put aside the nuclear fusion processes. Then a star can be considered as being thermally and kinematically in quasi-equilibrium. It slowly loses its energy by repeating radiation and contraction. The Virial theorem dictates that half of the gravitational energy produced by contraction is lost by radiation; but the other half is added to the internal energy, increasing the star's pressure, density, and temperature. When the pressure and the density reach those of a degenerate gas, the star can sustain pressure without extra energy supply. Consequently, the contraction stops and the star cools down by radiation but keeping its density. The drawing of Figure 2.51a describes the path of stellar evolution on the $\rho - T$ plane.

Depending on its total mass, which is denoted in units of the Chandrasekhar mass, a star enters different paths as indicated by real lines. The evolution begins as a protostar at the lower left of the lines and goes up the slope following the path dictated by $\ln T = (1/3) \ln \rho + \text{constant}$ [see Eq. (2.160)], where the constant depends on the total mass M. When the density reaches that of the degenerate gas, the contraction stops. Then, as the star loses its energy, the temperature goes

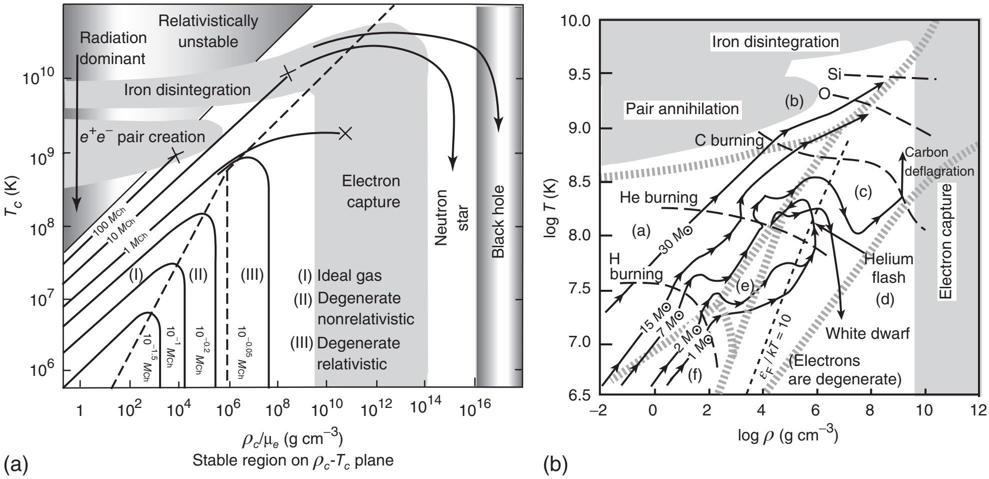


Figure 2.51 Stellar evolution paths. (a) Core density ρ_c versus core temperature T_c of a star having mass M. The mass is expressed in units of the Chandrasekhar mass $M_{\text{Ch}} \simeq 1.4M_{\odot}$ with $\mu_e = 2$. Shaded areas are unstable regions, upper left due to e^-e^+ pair creation or iron disintegration; the middle vertical band due to electron capture; and the rightmost band due to relativistic instability (black hole). The dashed lines separate nondegenerate electron region (I) from degenerate regions II (nonrelativistic) and III (relativistic). (b) Same as the left figure with inclusion of nuclear fusion effects. Starting from lower left, stars with different masses follow different paths, depending on whether the He/C burning ignition is in the nondegenerate or degenerate region. Also denoted are the major neutrino cooling processes separated by thin striped bands. They are (a) photoprocess ($\gamma + e^- \to e^- + \nu\overline{\nu}$), (b) pair-neutrino ($e^-e^+ \to \nu\overline{\nu}$), (c) plasma ($\gamma^* \to e^-e^+(\text{virtual}) \to \nu\overline{\nu}$), and (d) Bremsstrahlung ($e^- + Z \to e^- + \nu\overline{\nu}$). Stars that enter the unstable region explode either by deflagration or by core collapse. Figures adapted from [334].

down; that is, the path turns over and goes down on the straight line at constant density. If the total mass exceeds the Chandrasekhar mass, it keeps going up the slope until it reaches the unstable regions. As the temperature goes higher, one needs to consider radiation pressure, which had been neglected so far. It changes the equation of state and creates an unstable region denoted as *radiation dominant*, where the star cannot exist as a stable mass cluster.

So far we have neglected the chemical process and composition of the constituent particles. When the chemical composition changes, it changes the equation of state, and the path is no longer a straight line. Besides, electron–positron pair creation, electron capture, or iron disintegration creates another instability. In some cases, burning of newly created nuclei goes out of control and the star explodes, which is our next subject.

The Main Sequence (MS) When the temperature at the core exceeds 10^7 K, the hydrogen fusion process Eq. (2.62) begins:

$$4\text{H} + 2e^- = {}^4\text{He} + \nu_e + 26.73\ \text{MeV} \tag{2.163}$$

We refer to the process as *hydrogen burning*. Its fuel is hydrogen and its end product He is cinders. The ignition points of hydrogen (H) burning are denoted by a long–dashed line at the left bottom part of Figure 2.51b. Once the process commences, the energy loss from the surface is supplemented by the nuclear fusion and, as long as the fuel supply lasts, the gravitational contraction ceases and maintains stability. Stars in this state, which includes the Sun, are referred to as the *main sequence*. The Sun's core temperature is $\sim 1.8 \times 10^7$ K and p–p chain is the dominant process, but in stars at higher temperature $\gtrsim 2 \times 10^7$ K (i.e., heavier stars than the Sun), the CNO cycle becomes dominant (see Figure 2.23b). The fusion process is very sensitive to the temperature. Large stars (i.e., with mass several $M_\odot$) have higher pressure and higher temperature hence have higher burning rates. The fuel supply (i.e., mass) is larger but the burning rate is even faster, thereby leading to a faster death. Such stars have also higher surface temperature and shine in white-blue color. The hydrogen burning is a weak interaction process and its timescale is large.[38] Consequently, the process of hydrogen burning lasts for a long time. The Sun's original lifetime was about 10^{10} years, and its residual lifetime is $\sim 5 \times 10^9$ years. A star having $M \sim 10 M_\odot$ is calculated to have a lifetime of $\sim 10^7$ years.

Helium Burning As the hydrogen is consumed, helium accumulates at the core and an isothermal sphere of helium forms at the center. It is isothermal because the helium does not yet produce energy. The hydrogen is still burning at the surface of the helium core, keeping the pressure and density approximately constant. At the core, the helium is not ignited because of a high Coulomb barrier. This is the reason for the high ignition temperature of the helium burning. Hence the pressure and the temperature of the core go up. At this stage, the helium core, the hydrogen outer sphere, and their boundary evolve in different ways. As a result, a steep pressure gradient is formed, pushing the outer part of the hydrogen layer outward, and the star expands to become a red giant.

Eventually, the core pressure and temperature become high enough, helium fusion ($3\alpha \rightarrow ^{12}\text{C}$, $\alpha + ^{12}\text{C} \rightarrow ^{16}\text{O} + \gamma$) begins, and a core consisting of C + O begins to form. They are dense enough so that the electrons are in degenerate states. When the helium ignition occurs in the nondegenerate state, the burning raises the temperature and pressure with subsequent expansion, which, in turn, lowers the pressure and the temperature to settle in a new thermal equilibrium state. But if the star is already in a degenerate state, the pressure does not change, hence it does not expand and the self-cooling mechanism does not work. As the nuclear reaction rate is very sensitive to the temperature rise, the burning rate accelerates, the temperature goes up in a spiral, and a small explosion, referred to as the *helium flash*, ensues.

For the Sun or stars with similar mass, the helium ignition happens in a shallow degenerate state. When the temperature goes up, the star escapes from the degenerate region. After the flash subdues, the helium burns steadily, which is referred to as the *horizontal branch*. After the helium fuel exhaustion, if the temperature is not high enough to ignite the C + O core, the temperature rises again, the outer layer expands again, and becomes a bright red giant referred to as the *asymptotic giant branch*. As the outer layer is unstable, some part of it is blown out. The shining core and the blown-out clouds form what is referred to as the *planetary nebula*. The small degenerate core is referred to as the *white dwarf*. It keeps losing its energy by shining but keeps its density and fades away as a black dwarf. The Sun is believed to follow this pattern.

---

[38] If the process occurs through the strong interaction, it is a hydrogen bomb.

Carbon Deflagration If the star is not alone but has a companion, the outer layer is attracted by the companion and accretes on it, raising its temperature. When the temperature is high enough to ignite carbon-burning and the companion is already degenerate, the carbon-burning goes out of control and the so-called carbon deflagration ensues. The star explodes and leaves nothing behind. Stars with mass $8M_\odot < M < 12M_\odot$ are believed to fall in this category. The supernova in this form is referred to as *type Ia*. It is bright, and its luminosity variation as a function of time is well known. Type Ia supernovae are used as standard candles to measure cosmological distances.[39]

Fate of lone stars in this mass category will be discussed together with that of more massive stars.

Onion Structure of Presupernova Stars having mass $M > 8M_\odot$ and core mass larger than the Chandrasekhar mass proceed without falling into degenerate state, igniting more nuclear fuels and producing neon (Ne), magnesium (Mg), sulfur (S), silicon (Si), and so on, eventually reaching iron (Fe) if the star is heavy enough, that is, if $M > 12M_\odot$. Iron is the most stable nucleus with the largest binding energy (see Figure 2.52).

Therefore, the nucleosynthesis stops and the iron core develops in the center. The star is on the edge of its death. A progenitor of the supernova just before explosion has the onion structure with progressively heavier elements lying close to the center (see Figure 2.53 and Figure 2.54).

The energy liberated from hydrogen ignition to form the iron core is 8.6 MeV per nucleon. The hydrogen burning alone consumes 6.9 MeV of it and the burning of heavier elements occurs at much higher temperatures, and the residual life after hydrogen burnout is very short. The duration of hydrogen burning is $\sim 10^{10}$ years for the star with $M \sim M_\odot$, but helium burning lasts only for $\sim 10^7$ years.

Neutrino star: If the core temperature is higher than 10^8 K, the energy loss by neutrino emission (neutrino cooling) exceeds that of photon radiation from the surface. If more energy is emitted in the form of neutrinos, we might call it a *neutrino star*. The main neutrino production processes are described in Figure 2.55, and regions where they are the dominant cooler are denoted in Figure 2.51b.

---

[39] The expanding universe and the existence of the dark energy is discovered because the luminosity of supernovae Ia was well understood and was used to measure the absolute distance to them (see Chapter 11).

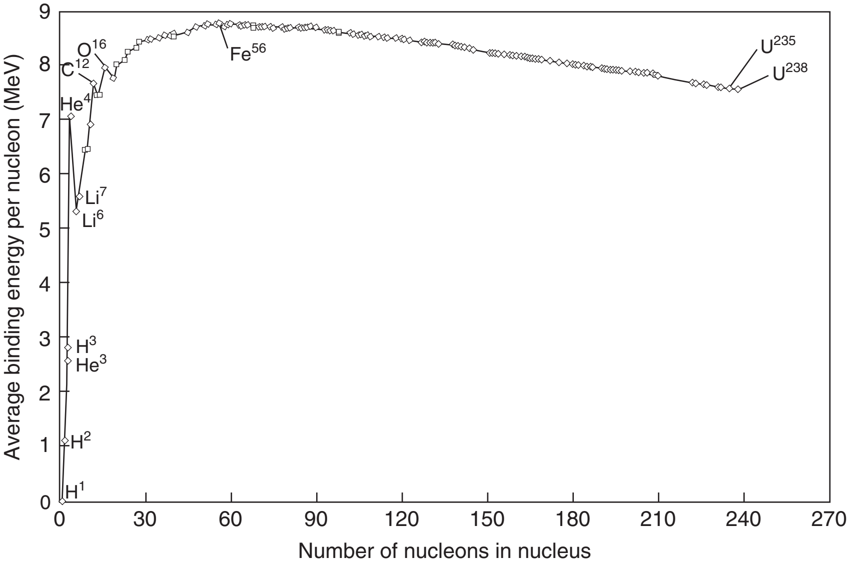


Figure 2.52 Nuclear binding energy per nucleon as a function of atomic number. Iron is the most stable, and no energy is produced by a process involving only irons. (Reproduced with permission of [335].)

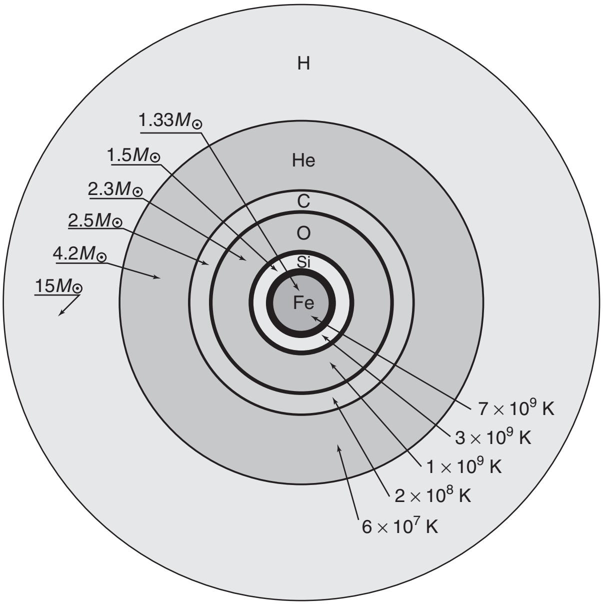

H
He
C
O
Si
Fe
1.33M⊙
1.5M⊙
2.3M⊙
2.5M⊙
4.2M⊙
15M⊙
7 × 10^9 K
3 × 10^9 K
1 × 10^9 K
2 × 10^8 K
6 × 10^7 K

Figure 2.53 Illustration showing the onion structure of a star just before explosion (scale is arbitrary).

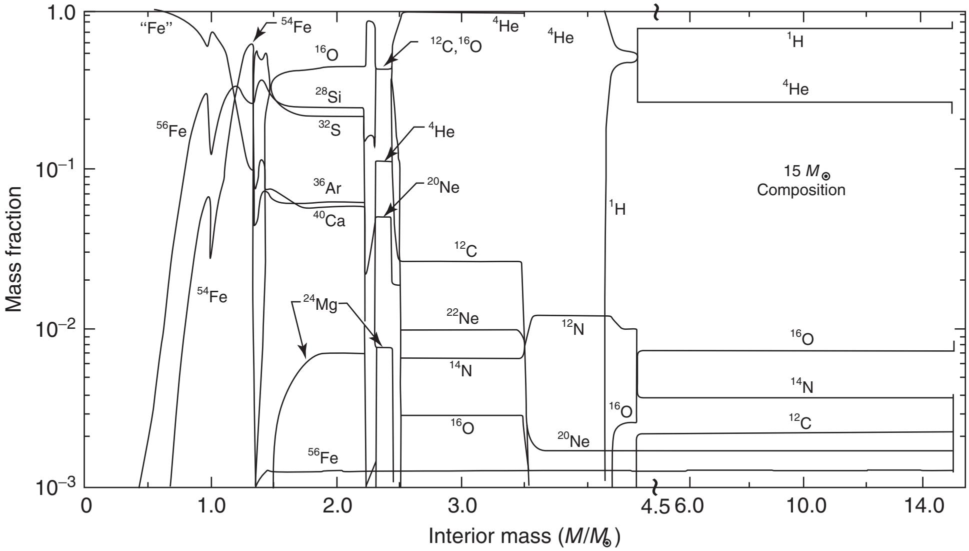


Figure 2.54 Picture showing an onion structure and composition of a 15$M_\odot$ presupernova star at a time when the edge of its iron core begins collapsing at 1000 km s^{-1}. Neutrino emission from electron capture dominates photodisintegration in the total energy losses throughout most of the iron core. Central temperature here is 7.62×10^9 K and density is 9.95×10^9 g cm^{-3}. Note a scale break at 5$M_\odot$. (Reproduced with permission of [336, 337].)

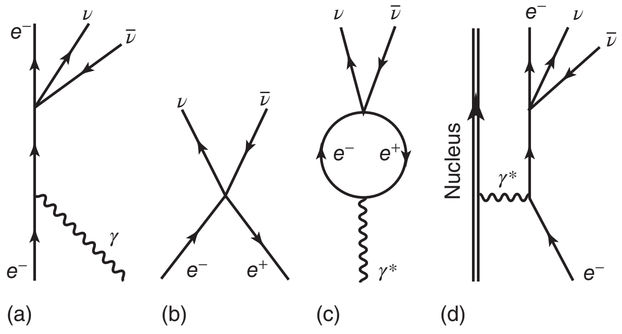


Figure 2.55 Neutrino cooling reactions in the interior of aged stars. (a) Photoneutrino process. (b) Pair-neutrino process. (c) Plasmon decay. (d) Bremsstrahlung. Regions on the ρ–T plane where each process becomes dominant are denoted in Figure 2.51b.

The star reaches this stage at the time of carbon burning. Once the neutrino cooling turns on, the fuel consumption rate accelerates because the neutrino takes out energy directly from the core, making itself a very efficient cooler. The burning duration of the oxygen is $\sim 10^2$ years, and that of silicon $\sim$ a month. The timescales of the post-silicon processes are even shorter.

Core Collapse: Death of a Star As the nuclear burning shell makes the stellar density distribution nonuniform, the mass in Figure 2.51 should be considered as the core mass rather than the total mass. The core mass is considerably smaller than the total mass; only stars with total mass $M > 8M_\odot$ make a core larger than the Chandrasekhar mass. In order to reach iron or Si ignition, the temperature has to be higher than $\sim 10^{9.6}$ K and yet be nondegenerate when Si starts burning. Only stars with mass $M > 12M_\odot$ can reach this stage. Stars with $8M_\odot < M < 12M_\odot$ have properties intermediate between small stars, which end up as white dwarfs, and massive stars, which go into core collapse. Depending on the size of the core mass, some fall into a degenerate state, ending their life as white dwarfs, but others ignite the carbon in the nondegenerate state, further raising the temperature and pressure. Although nondegenerate, the core comes near to degeneracy and the electrons near the Fermi level have very high energy $\varepsilon \sim m_e$ and induce cascade electron capture processes. The nuclei are neutralized and the number of electrons is reduced, reducing the pressure and eventually running into the violent gravitational collapse.

Heavier stars with $M > 12M_\odot$ reach the iron-forming stage, keeping its nondegenerate state. As the iron does not produce energy any more, the core begins to contract again, raising its pressure and temperature. Eventually, photodisintegration of the iron core commences.

$$\gamma + {}^{56}\mathrm{Fe} \to 13\,{}^4\mathrm{He} + 4n - 124.4\,\mathrm{MeV} \tag{2.164a}$$

$$\gamma + {}^4\mathrm{He} \to 2p + 2n - 28.3\,\mathrm{MeV} \tag{2.164b}$$

This is an endothermic reaction and absorbs energy, reducing the pressure and accelerating the contraction. The disintegration process happens in ~ 0.1 s. About the same amount of energy consumed so far spending $\sim 10^{8-10}$ years via nuclear fusion is absorbed, the core supporting pressure suddenly vanishes and violent fall into the center referred to as *implosion* ensues. Since the mass of the collapsing core exceeds the Chandrasekhar limit, the degenerate electrons cannot withstand the pressure of the collapsing core. Protons and electrons fuse to convert to neutrons (neutronization: $e^- + p \to \nu_e + n$).

Type II Supernova The falling core contracts uniformly (homologous collapse). When the central density exceeds the nuclear density ($\sim 3 \times 10^{14}\,\mathrm{g\,cm^{-3}}$), the nucleons that are mostly neutrons due to neutronization contact each other without gaps. The strong repulsive nuclear force becomes effective, and a hard core is formed. The gravitational collapse ceases at a few times the nuclear density.

At this stage, the neutrons are degenerate and sustain the core's finite size by their degenerate pressure, which has again an upper limit that is not much larger than the Chandrasekhar mass. If the growing core mass exceeds its limit, the degenerate pressure can no longer sustain the falling nuclei and the core becomes a black hole.

The falling materials from outside the core collide with the hard core and bounce. A shock wave is produced that propagates outward. If it reaches the surface, it becomes a supernova (prompt explosion). However, it is known that the expanding waves stall at a radius of ~200 km. The reason is that the shock front produces a high-pressure, high-temperature layer in which the falling iron and helium disintegrate, absorbing the shock energy.

Neutrinosphere: At this time, a large number of neutrinos are produced inside the neutron core [Eq. (2.165)]. The high density of the core makes the mean free path short. It becomes opaque even to the neutrinos and, as a consequence, the neutrinos are thermalized. The weak NC (neutral current) interaction acting coherently on nuclei, whose strength is proportional to A^2, enhances the acceleration to make the nuclear material opaque to neutrinos. The boundary where it becomes transparent to the neutrino is referred to as the *neutrinosphere*. It is a concept similar to the photosphere of the Sun.

The thermal neutrinos emitted from the neutrinosphere and stored behind the shock wave produce high pressure, pushing back the stalling shock front, and revive the outgoing shock wave (delayed explosion) (see Figure 2.56a). Note, however, that

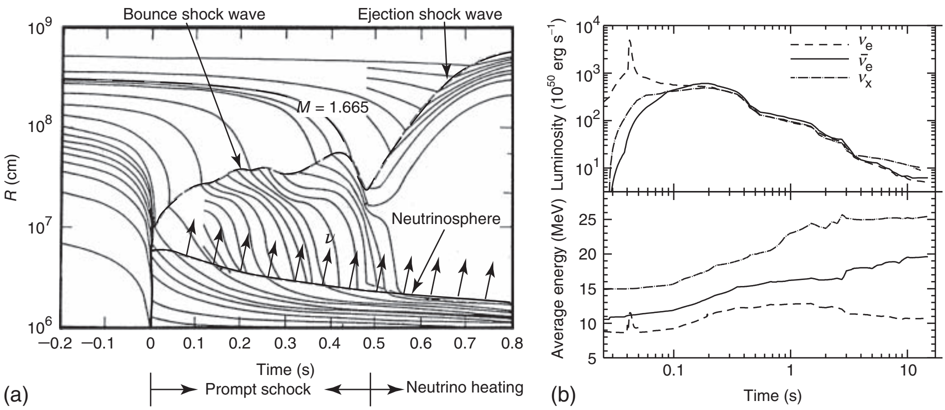


Figure 2.56 (a) Successful delayed explosion. Abscissa and ordinate represent the time (in units of seconds) measured from core bounce and the radius from the stellar center (in units of centimeters), respectively. Lines are trajectories of selected mass zones. The dashed line represents the shock front. In (a), 1.665*M* shows the mass point that is expelled outward by the second shock produced by the neutrino heating of the matter behind the stalled shock wave. Because of the neutrino heating mechanism, the shock wave once weakened at ~500 ms revives and then successfully explodes. [338, 339], and propagates to the surface of the iron core. Figure adapted from [338, 340, 341] (b) Time evolution of neutrino luminosity and average energy of the numerical supernova model. The dashed line is for ν_e, solid line for $\bar{\nu}_e$, and dot-dashed line for ν_x (each of ν_μ, ν_τ, $\bar{\nu}_\mu$, and $\bar{\nu}_\tau$). The core bounce time is 3 − 4 ms before the neutronization burst of ν_e's. (Reproduced with permission of [342].)

the details of the explosion mechanism are not well understood yet. Simulations of supernova explosions are not always successful. Which of the mechanisms (prompt or delayed) is effective is an unsettled problem and is a subject of intense research. Studies on the effects due to convection and magnetic fields are still immature.

Supernova Neutrino After the carbon burning, the energy loss due to the neutrino dominates that of photons and, at the core collapse stage, more than 99% of the liberated energy is carried by the neutrinos. At the beginning of the core collapse, a large number of neutrinos are produced by neutronization ($e^- p \to \nu_e n$). They last only for a few milliseconds, go out promptly, and are observed as a sharp peak shown in Figure 2.56b. In the mean time, the core density keeps increasing, and becomes opaque even to the neutrinos. Neutrinos produced inside the core are thermalized by multiple scattering, diffuse and highly degenerate, and escape the core only from the surface (neutrinosphere outside of which becomes transparent to the neutrinos). When the neutrinosphere is formed, it has an initial radius $\sim$1000 km. Inside it, neutrinos are produced by thermal processes:

$$\begin{aligned} \gamma + \gamma &\to \nu_e\overline{\nu}_e,\ \nu_\mu\overline{\nu}_\mu,\ \nu_\tau\overline{\nu}_\tau \\ e^- + e^+ &\to \nu_e\overline{\nu}_e,\ \nu_\mu\overline{\nu}_\mu,\ \nu_\tau\overline{\nu}_\tau \end{aligned} \tag{2.165}$$

The emitted neutrinos from the surface deprive energy out of the neutrinosphere. When the core eventually settles as a neutron star, the radius shrinks to $\sim$10 km. All flavors of neutrinos are produced more or less in equal numbers, and the thermal neutrino emission lasts for several seconds (Figure 2.56b). Since ν_μ, ν_τ have no CC (charge current) interactions, they have lower opacity and decouple at high temperature and density. The opacity for $\overline{\nu}_e$ is also lower because fewer protons are available than neutrons. Therefore, the average energies are typically $< E(\nu_e) >= 10 - 12$ MeV, $< E(\overline{\nu}_e) >= 14 - 17$ MeV, and $< E(\nu_\mu, \nu_\tau) >= 24 - 27$ MeV. The neutrino spectrum might well be described by the degenerate Fermi–Dirac black body formula. Note that the neutrino emission is about 3 h before optical explosion, which is the time taken by the shock wave to reach the surface of the star.

The supernova neutrinos can be observed terrestrially by the following processes [343]:

$$\begin{aligned} \sigma(\overline{\nu}_e p \to e^+ n) &= 9.77 \times 10^{-42}\,(E_\nu/10\ \text{MeV})^2\ \text{cm}^2 \\ \sigma(\overline{\nu}_e e \to \overline{\nu}_e e) &= 0.388 \times 10^{-43}\,(E_\nu/10\ \text{MeV})\ \text{cm}^2 \\ \sigma(\nu_e n \to e^- p) &= 2.2 \times 10^{-42}\,(E_\nu/10\ \text{MeV})^2\ \text{cm}^2 \\ \sigma(\nu_e e \to \nu_e e) &= 0.933 \times 10^{-43}\,(E_\nu/10\ \text{MeV})\ \text{cm}^2 \\ \sigma(\nu_\mu e \to \nu_\mu e) &= 0.159 \times 10^{-43}\,(E_\nu/10\ \text{MeV})\ \text{cm}^2 \\ \sigma(\overline{\nu}_\mu e \to \overline{\nu}_\mu e) &= 0.130 \times 10^{-43}\,(E_\nu/10\ \text{MeV})\ \text{cm}^2 \\ \sigma(\nu_\tau e \to \nu_\tau e) &= \sigma(\nu_\mu e \to \nu_\mu e) \\ \sigma(\overline{\nu}_\tau e \to \overline{\nu}_\tau e) &= \sigma(\overline{\nu}_\mu e \to \overline{\nu}_\mu e) \\ \sigma(\nu_e^{16}\text{O} \to e^{16}\text{F}) &= 1.1 \times 10^{-44}\,(E_\nu - 13\ \text{MeV})^2\ \text{cm}^2 \end{aligned} \tag{2.166}$$

If one uses water to detect the neutrinos, there are no free neutrons and the important process is $\overline{\nu}_e p \to e^+ n$. In this process, the electrons are scattered isotropically and information on the incident flight direction is lost. The $\nu_e e$ process, although less frequent $(N(\overline{\nu}_e p)/N(\nu_e e) \simeq 20/1)$, is important because it can point to the incoming direction. Despite their large masses (2140 t for Kamiokande and 6800 t for IMB), the number of observed events is small, 11 for Kamiokande and 8 for IMB. They are presumably all $\overline{\nu}_e$ events.

Neutrino Luminosity Assuming that the energy spectrum of the neutrino is that of degenerate Fermi–Dirac statistics and taking into account the detector efficiency and distance to the SN1987A (50 kpc), one can calculate the neutrino temperature T_ν and total energy $L(\overline{\nu}_e)$ from the observed number and energy of events [344].

$$T_\nu = 3.7 \pm 0.4\,\mathrm{MeV}, \qquad L(\overline{\nu}_e) = 4.8 \pm 2 \times 10^{52}\,\mathrm{erg} \tag{2.167}$$

Multiplying by 6 to take into account the other flavors, the total energy emitted by the SN1987A is calculated to be

$$E = 3 \pm 1 \times 10^{53}\,\mathrm{erg} \tag{2.168}$$

This number is to be compared with the theoretical prediction, which can be derived assuming a uniform density distribution. Attaching suffixes "star" and "ns" to the mass M and radius R of the star before the explosion and the residual neutron star, one can express the liberated energy as

$$E = \left(-\frac{5}{3} G_N \frac{M_{\mathrm{star}}^2}{R_{\mathrm{star}}} \right) - \left(-\frac{5}{3} G_N \frac{M_{\mathrm{ns}}^2}{R_{\mathrm{ns}}} \right) \tag{2.169}$$

where G_N is the gravitational constant. With $R_{\mathrm{star}} \sim 10^{10}$ cm and $R_{\mathrm{ns}} \sim 10^6$ cm, the second term is dominant despite the fact $M_{\mathrm{star}} \gg M_{\mathrm{ns}}$, and the total energy is given by [345]

$$E \simeq 2.5 \pm 1.5 \times 10^{53}\,\mathrm{erg} \left(\frac{R_{\mathrm{ns}}}{10\,\mathrm{km}} \right) \left(\frac{M_{\mathrm{ns}}}{1.4 M_\odot} \right)^2 \tag{2.170}$$

$M_\odot = 1.989 \times 10^{30}$ kg is the solar mass. The agreement with the observation is excellent.

Stellar Cycle Thus, heavy stars end their lives gloriously. But this is the beginning of other stars. Stars are created from the ISM in the form of giant clouds. Gravitational attraction creates nonuniformity of the mass distribution, which attracts more materials, eventually forming stars. Depending on the initial mass, many, including our Sun, end up as white dwarfs after exhausting hydrogen fuels, but some go all the way to the iron to become supernovae or black holes entering another stellar cycle (see Figure 2.57). Formation of the solar system is believed to have been triggered by the explosion of a nearby supernova. Its shock wave created local dense clusters of gas clouds, which evolved into the solar system.

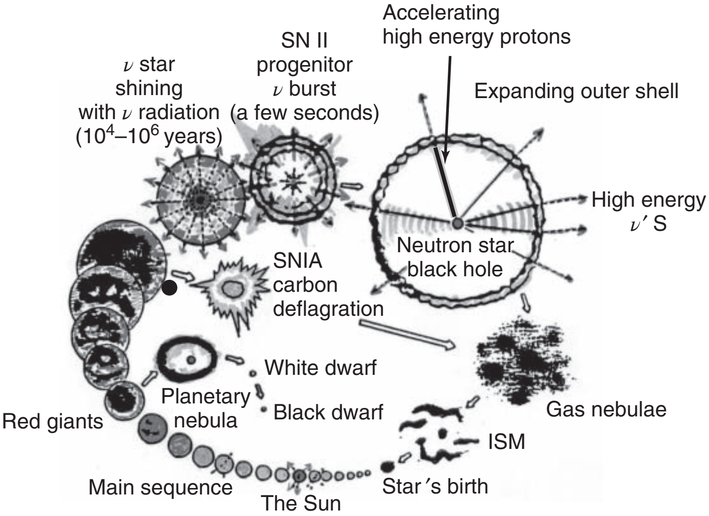

Accelerating
high energy protons
SN II
progenitor
ν burst
(a few seconds)
ν star
shining
with ν radiation
(10^4–10^6 years)
Expanding outer shell
High energy
ν′ S
Neutron star
black hole
SNIA
carbon
deflagration
White dwarf
Planetary
nebula
Black dwarf
Gas nebulae
Red giants
ISM
Main sequence
The Sun
Star′s birth

Figure 2.57 Stellar cycle. Life of stars.

In the beginning, soon after the big bang, only light elements (H, D, Li, B, etc.) existed (see Section 8.6.5 on nucleosynthesis). Hydrogen occupied three-fourths of all the particles in the universe and helium one-fourth, and the rest of light elements occupied tiny fractions. The first-generation stars are thus composed almost exclusively of hydrogen and helium. Elements heavier than carbon are synthesized inside the stars. Thus stars that contain heavy elements including the Sun belong to the second or later generations.

2.10.2
Feedback to Particle Physics

Since known interactions alone can explain the supernova energy, there is no space for new interactions to come in. Using this fact, one can constrain the properties of yet-to-be-discovered new particles. The magnetic moment of the neutrino discussed in Section 2.3 [see Eq. (2.42)] is based on this fact. Below, we extract several new pieces of information on particle physics [214, 346].

Mass of the Neutrino If the neutrino is massive, its velocity is less than the speed of light c [347, 348, 349].

$$v_\nu \simeq 1 - \frac{m_\nu^2}{2E_\nu^2} \tag{2.171}$$

Denoting the distance to SN1987A as L, the extra time that the neutrino takes to reach the earth is given by

$$\Delta t = 2.57\,\mathrm{s} \times \left(\frac{L}{50\,\mathrm{kpc}}\right)\left(\frac{10\,\mathrm{MeV}}{E_\nu}\right)^2\left(\frac{m_\nu}{10\,\mathrm{eV}}\right)^2 \tag{2.172}$$

The time difference between neutrinos having E_1 and E_2 becomes

$$\Delta t_{\text{mass}} = 2.57\,\text{s} \times \left(\frac{L}{50\,\text{kpc}}\right)\left(\frac{m_\nu}{10\,\text{eV}}\right)^2 \left[\left(\frac{10\,\text{MeV}}{E_1}\right)^2 - \left(\frac{10\,\text{MeV}}{E_2}\right)^2\right] \tag{2.173}$$

Therefore, by looking at the time distribution of the arriving neutrinos, one can estimate their mass. All the observed neutrinos are within interval of 10 s and constrain the mass to $m_\nu \lesssim 20$ eV. A recent analysis gives $m_\nu < 5.7$eV [350]. For the ν_μ and ν_τ, this gives much better results than laboratory measurements although not as good as the cosmological limit of $\sum_i m_{\nu_i} < 0.14$ eV [see Eq. (2.2)] [119, 122, 121].

Life Time of the Neutrino The fact that the neutrinos from SN1987A were observed means that they did not decay in their flight. Therefore, considering the relativistic time dilation

$$\left(\frac{E_\nu}{m_\nu}\right)\tau_\nu \gtrsim 5 \times 10^{12}\,\text{s} \quad \rightarrow \quad \tau_\nu \left(\frac{E_\nu}{10\,\text{MeV}}\right) \geq 5 \times 10^5 \left(\frac{m_\nu}{1\,\text{eV}}\right)\,\text{s} \tag{2.174}$$

Electric Charge of the Neutrino If the neutrino carries an electric charge, its flight path would be deflected by the galactic magnetic field ($B \sim 10^{-6}$ G) and its arrival time on Earth would take more time than without the charge. Denoting L_B as the straight pathlength where the magnetic field exists, the time difference would be [351]

$$\Delta t_Q = \frac{L_B^3 Q_\nu^2 e^2 B^2}{24 m_\nu^2}\left(1 + \frac{m_\nu^2}{E_\nu^2}\right) \tag{2.175}$$

which enlarges the time dispersion due to finite mass effect

$$\Delta t = \Delta t_{\text{mass}} + \frac{L_B^3 Q_\nu^2 e^2 B^2}{24}\left(\frac{1}{E_1^2} - \frac{1}{E_2^2}\right) \tag{2.176}$$

The upper limit of the electric charge can be rewritten using the mass upper limit and gives

$$Q_\nu < 3 \times 10^{-17}\left(\frac{10^{-6}\,\text{G}}{B}\right)\left(\frac{1\,\text{kpc}}{L_B}\right) \tag{2.177}$$

Magnetic Moment of the Neutrino If the neutrino has magnetic moment $\kappa\mu_{Bohr}$, the right-handed neutrinos ν_R's would be produced in the neutrinosphere of the supernova. Then they interact only through the magnetic interaction or through coupling with the right-handed W boson with weaker interactions and tend to exit directly and reduce the neutrino luminosity. Requiring the reduction does not affect the observed luminosity; one can constrain the value of κ [158, 352, 353].

Another effect comes from the fact that they exit the neutrinosphere directly, that is, without being thermalized. Therefore, they have high energies ($E_\nu = 100 - 200$ MeV) compared to other neutrinos. If they carry a magnetic moment, they can go back to left-handed neutrinos through interactions with the galactic magnetic field. As no high-energy neutrinos were observed, one can constrain κ to

$$\kappa \lesssim 10^{-13} \tag{2.178}$$

The limit is considerably stronger than that obtained from laboratory experiments and from the stellar cooling arguments (see paragraph "Magnetic moment of the neutrino" in Section 2.3). However, one should note that the supernova arguments rely on the not-yet-well-understood neutrino formation mechanism inside the supernova.

Oscillation in the Supernova As the density in the supernova core is high, there are two level-crossings, as shown in Figure 2.58 [354, 355, 356, 357]. There are no level-crossings for the anti-neutrinos in the NH. Hence, the $\overline{\nu}_e$ produced in the supernova will come out unchanged and reach the Earth. On the other hand, for the IH, a level-crossing for $\overline{\nu}_e$ occurs at the higher density, as shown in Figure 2.58b. If there is a large mixing, the high-energy $\overline{\nu}_\mu$, $\overline{\nu}_\tau$ would be converted to high-energy $\overline{\nu}_e$ and the low-energy $\overline{\nu}_e$ produced in the core would be converted to low-energy $\overline{\nu}_\mu$, $\overline{\nu}_\tau$ when they come out. As no high-energy $\overline{\nu}_e$ was observed and the observed number of events is consistent with the standard supernova formation theory, qualitatively the IH may be disfavored [358]. However, quantitatively it is hard to draw a definite conclusion.

It is also argued that oscillations in the channel $\nu_e - \nu_x$ $(x = \mu, \tau)$ are able to reduce the prompt neutrino significantly. It is argued that, if the oscillation resonances occur outside the neutrinosphere but inside the shock wave (which is valid if

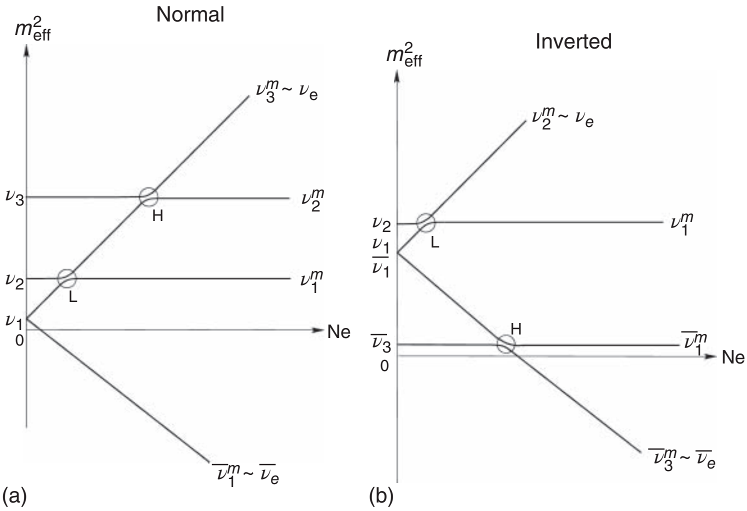


Figure 2.58 Schematic level-crossing diagram for the case of (a) normal and (b) inverted mass hierarchies. The circles with the symbol H and L correspond to resonances that occur at higher and lower density, respectively. (Reproduced with permission of [358].)

$m_\nu \gtrsim 10$ eV), the higher energy ν_μ, ν_τ can be converted to ν_e to revive the stalled shock [359]. An interesting byproduct is the change of r-process nucleosynthesis rate.[40] The chemical composition is affected by the ratio p/n in the hot bubble between the settled neutron star and the expanding shock wave which, in turn, is governed by the reactions $\nu_e n \leftrightarrow e^- p$ and $\overline{\nu}_e p \leftrightarrow e^+ n$. Under normal circumstances, $\overline{\nu}_e$ is more energetic, and hence more neutrons are produced. The oscillation could enhance the ν_e energy and drive the medium to a proton-rich state [360].

---

[40] r-process is a succession of rapid nucleosynthesis by neutron capture which occurs behind the fast expanding shock wave of the supernova starting typically from ^{56}Ni and synthesizes all kinds of heavier elements. There is another mechanism to make heavy elements, referred to as the *s-process*, which is similar but is induced by slow neutrons whose capture rate is slower than typical beta-decay rate and mainly happens in red giants (AGB, asymptotic giant branch). Cosmic chemical abundance of heavy elements (heavier than iron) are explained by the r- and s- processes, each sharing about half of the abundance.

3
Grand Unified Theories

3.1
Introduction

The Standard Model (SM) is almost infallible except perhaps for the neutrino mass problem mentioned in the previous chapter. Theoretically, however, there are many shortcomings and many model extensions are proposed. Here we try to investigate its weak points in general, not just the neutrino problem, to find ways to solve them and possibly to unite the three interactions (referred to as grand unification), namely the weak, the electromagnetic, and the strong, in one package. Consequently, the following discussions necessarily become more speculative. However, detailed discussions of models without plausible justifications are time consuming and fruitless. In the following, we introduce a grand unified theory (GUT) based on $SU(5)$ [361, 362], which is a direct extension of the SM, discuss its major weak point, that is, the hierarchy problem. Possible ways to solve it are discussed, and the concept of the supersymmetry (SUSY) is introduced as the most promising remedy [363]. SUSY is an interesting theme of its own, especially because, by gauging the symmetry, one can naturally include the gravity within the same mathematical framework with other interactions, and it is thus considered as an essential ingredient to reach the ultimate unification. We also discuss a few topics in the GUTs that are testable.

3.2
Why GUTs?

Let us start listing the parameters of the SM that cannot be determined from the first principle (see Table 3.1).

Here, we did not include the neutrino mass and the PMNS matrix. Compared to Einstein's equation of gravity which contains only one parameter, namely the gravitational constant, it is obvious that the SM is far from a unified theory. Especially important among those is the electroweak symmetry-breaking mechanism, that is, the mass hierarchy problem that was discussed in Chapter 1. We also argued in the previous chapter that, in order to understand the neutrino mass problem, the SM needs to be extended.

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

How many quarks and leptons are there? Why is the generation repeated three and only three times (see Figure 3.1).

This is the generation problem or, more generally, the flavor problem. This question was first raised by Rabi about the role of the muon, "who ordered it?" in as early as the 1930s. It was a question of having a redundant heavy electron but the question is enlarged to all the quarks and leptons beyond the first generation. It remains an unanswered question to this day. One possibility is that the quarks and leptons are composites of more fundamental particles (subquarks) and that higher generations are excited states. However, whether particles of mass differing as much as $m_t/m_u \sim 10^5$ across the generations or $m_t/m_\nu > 10^{12}$ within a generation

Table 3.1 Parameters in the Standard Model.

Coupling constants	G_F, α, α_s	
Masses	Gauge bosons;	$W^\pm$, Z^0
	Fermions;	Six quarks and three leptons
	Scalar particles	Higgs
CKM matrix	$\theta_{12}, \theta_{23}, \theta_{13}, \delta$	
Number of generations	$n_g = 3$	
CP parameter in QCD	θ	

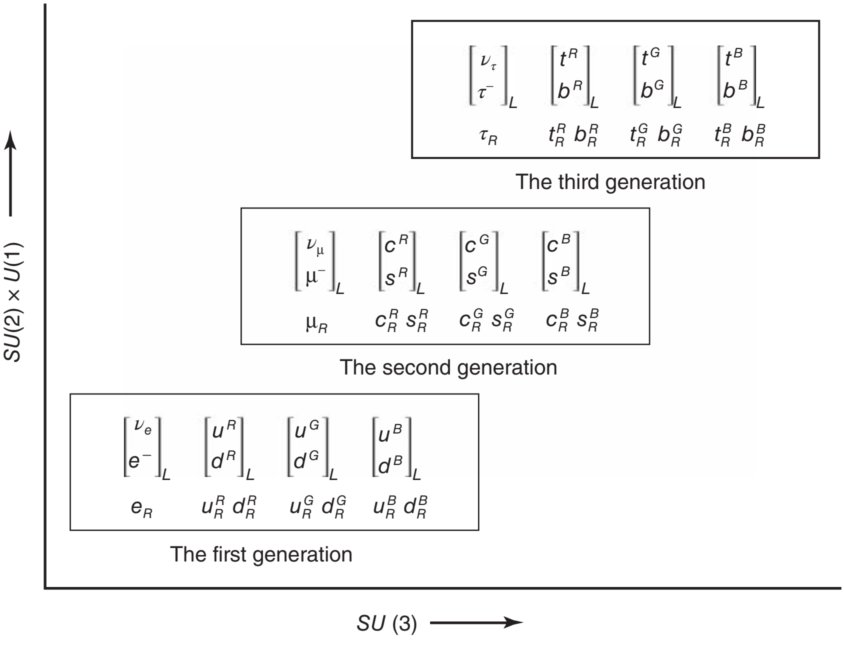

$\begin{bmatrix}\nu_\tau\\ \tau^-\end{bmatrix}_L$ $\begin{bmatrix}t^R\\ b^R\end{bmatrix}_L$ $\begin{bmatrix}t^G\\ b^G\end{bmatrix}_L$ $\begin{bmatrix}t^B\\ b^B\end{bmatrix}_L$
τ_R $t^R_R\ b^R_R$ $t^G_R\ b^G_R$ $t^B_R\ b^B_R$
The third generation
$\begin{bmatrix}\nu_\mu\\ \mu^-\end{bmatrix}_L$ $\begin{bmatrix}c^R\\ s^R\end{bmatrix}_L$ $\begin{bmatrix}c^G\\ s^G\end{bmatrix}_L$ $\begin{bmatrix}c^B\\ s^B\end{bmatrix}_L$
μ_R $c^R_R\ s^R_R$ $c^G_R\ s^G_R$ $c^B_R\ s^B_R$
The second generation
$\begin{bmatrix}\nu_e\\ e^-\end{bmatrix}_L$ $\begin{bmatrix}u^R\\ d^R\end{bmatrix}_L$ $\begin{bmatrix}u^G\\ d^G\end{bmatrix}_L$ $\begin{bmatrix}u^B\\ d^B\end{bmatrix}_L$
e_R $u^R_R\ d^R_R$ $u^G_R\ d^G_R$ $u^B_R\ d^B_R$
The first generation
$SU(2) \times U(1)$
$SU(3)$

Figure 3.1 Families of particles and generation structure. Each generation consists of doublets of left-handed fields and singlets of right-handed fields in the electroweak interaction. In addition, the quarks also belong to triplets in QCD or carry three colors denoted by their upper suffix R, G, B. The right-handed neutrino is absent in the Standard Model.

can be treated on the same footing is a big question. In the following, we list a few items that can be solved by going to GUTs.

3.2.1
Weinberg Angle in GUTs

The SM is described in the framework of the gauge theory based on $SU(3)_{\text{color}} \times SU(2)_L \times U(1)_Y$ symmetry. The electromagnetic and the weak force are partially unified. By "partially unified" we mean that the groups $SU(2)$ and $U(1)$ are not subgroups of a single larger group, and therefore the coupling constant of $SU(2)$, g_W, and $U(1)$ g_B are independent. Consequently, the Weinberg angle, which is defined as their ratio, is also arbitrary. Naturally, our next target is to find a larger framework in which the Weinberg angle can be derived from the first principle.

QCD, on the other hand, seems detached from the electroweak interaction, but considering that its mathematical framework is also the gauge theory, we are naturally led to a conjecture that all the three interactions may be unified under a single symmetry group. If we can find it, the three coupling constants, including the Weinberg angle, will be reduced to a single parameter. The GUT is a name given to a theory to unify the three interactions. If one includes gravity also, one may call it the *super GUT*. However, there are no accepted names for it. Models that can be constructed by gauging the SUSY are given a special name, supergravity. Currently, most popular string theories consider that the fundamental constituent of matter is not a particle but a string. They also incorporate many new ideas including the SUSY and extra dimension, which we will discuss later.

3.2.2
Quantization of the Electric Charge

We have more reasons to believe in the existence of a larger group. In the SM, the electromagnetic interaction belongs to the $U(1)$ gauge group. The group $U(1)$ is Abelian, its equation of field is linear, and normalization of the coupling constant is arbitrary. In other words, the electric charge has no compelling reason to be quantized. However, we know that the magnitude of the electric charges of the proton and electron agrees with each other to within an accuracy of 10^{-21} [364] and that of the quark is one-third of it. These facts strongly suggest that the quantization of the charge is a built-in structure of particle theories. It can be achieved by requiring that the electric charge is a conserved quantity made of group generators of a simple, large group. We write the generators of the larger group G as L_i. They satisfy the traceless condition.

$$\mathrm{Tr}[L_i] = 0^{1)} \tag{3.1}$$

---

1) The requirement of the absence of the triangular anomaly achieves the same purpose. Some people argue that it is more fundamental.

The relation holds in any $SU(N)$ groups for $N > 1$ (see Appendix G of [1]). As the group G has to contain the SM, we have a constraint

$$G \supset SU(3) \times SU(2) \times U(1) \tag{3.2}$$

3.2.3
Triangle Anomaly

We know that leptons do not interact strongly and that they belong to a distinct class different from quarks. Although they have electric charges in common, mixing between the lepton and the quark has not been observed. The absence of mixing is guaranteed if the lepton number (N_L) and/or quark number (N_q) are conserved quantities.[2] Note, however, that conservation of the lepton number and the quark number is purely phenomenological and is not based on any symmetry.

Within the framework of unified gauge theories, conservation laws that are not based on the local gauge invariance cause no trouble if they are actually violated. Once we accept the notion that the lepton or the quark number need not be conserved, a whole new world is open. Let us recollect the family structure of the quarks and leptons. A family is composed of two quarks (or six including color degrees of freedom) and two leptons. Furthermore, the theory of the SM can be closed within a family (or generation) (see Figure 3.1). If quarks and leptons are part of a family, they can mix. If they do, there should be a new interaction that violates the lepton and/or quark number conservation. This is a drastic conjecture because phenomenologically there are no evidences for its arguments. Theoretically, however, there is a strong reason to consider the family structure, a requirement to eliminate the triangular anomaly [365, 366].

The axial current $J_5^\mu = g\overline{\psi}\gamma^\mu\gamma^5\psi$ made of fermions belonging to some gauge group G conserves at the tree level if they are massless. However, when one includes quantum corrections made of triangle loops like those depicted in Figure 3.2, the current is no longer conserved.

This is referred to as *the quantum (or chiral or triangle or axial or Adler–Bell–Jackiew) anomaly* which violates the symmetry in a way different from external or spontaneous symmetry breaking. The former means that an explicit symmetry-breaking environment is provided, or, theoretically, a symmetry-breaking term is added to the Lagrangian. The latter happens when the solution to the equation of motion respects the original symmetry but is unstable. The quantum anomaly arises when the symmetry is respected at the classical Lagrangian level but is broken when quantum effects are taken into account. It happens when one or three of the external lines in Figure 3.2b couple to the axial current and the rest to the polar vector current.[3] When one includes the contribution of the triangular processes, the divergence of the current is

---

[2] Some people use the baryon number for historical reasons. Strictly speaking, $N_B = 3N_q$, but many people use the baryon number to mean the quark number as well.

[3] There are other cases to induce the quantum anomaly, but they always reduce to the triangle anomaly described here.

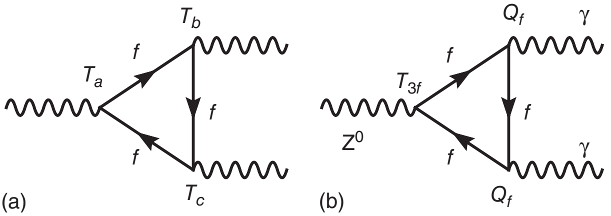


Figure 3.2 Triangular anomaly. Three currents couple to a fermion loop. (a) A general coupling type. The current couples to the fermion through gauge generators t_a, t_b, t_c. (b) Coupling of $Z^0 - \gamma - \gamma$. Z^0 couples via t_{3f} and the photons through charge Q_f.

expressed as

$$\partial_\mu J_5^\mu = \frac{g^2}{32\pi^2}\tilde{F}^{\mu\nu}F_{\mu\nu} = \frac{g^2}{64\pi^2}\varepsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} \tag{3.3}$$

where $F_{\mu\nu}$'s are the field strengths that couple to the fermion fields and here we have considered only the Abelian group. For the non-Abelian group, a coefficient proportional to

$$2\mathrm{Tr}\Big[t_a\{t_b,\ t_c\}\Big], \qquad \{A,\ B\} = AB + BC \tag{3.4}$$

is multiplied to the current where t_a's are generators of the symmetry group. Chiral symmetries that distinguish left- and right-handed currents always contain this axial anomaly. If the symmetry is global, there is no harm to include this contribution. In fact, it is known that the decay $\pi^0 \to 2\gamma$ is well explained by the anomaly term [see Eq. (14.42) in [1]]. However, if the symmetry is local, that is, if the current is a gauge current, the conservation law is broken by the existence of the axial anomaly and, as a consequence, the renormalizability is broken. This is serious because the theoretical framework of dynamics is no longer justified.

The electroweak interaction is based on chiral $SU(2)_L \times U(1)_Y$ symmetry and generation of the axial anomaly is unavoidable. If the renormalizability is a necessary condition for justifiable theories, it must not exist. A condition for the overall anomaly to vanish is that sum of the anomaly contributions cancel each other. As a simple example, let us consider $Z - \gamma\gamma$ decay shown in Figure 3.2b. In the SM, the coupling of Z to the fermion (neutral current) is proportional to $I_3 + Q\sin^2\theta_W$, but only the first term contributes to the axial anomaly. The other two currents are due to photons which are polar vectors. Then, according to Eq. (3.4)

$$\text{Axial anomaly} \quad \propto g_Z^2 \sum_f \mathrm{Tr}\left[I_{3f} Q_f^2 N_f\right] = 0 \tag{3.5}$$

and the sum of the anomaly contributions would be canceled. Here, N_f is the color factor, which is 3 for quarks and 1 for leptons. The argument goes as follows. As

$Q = I_3 + Y/2$ and $\mathrm{Tr}[I_3] = \mathrm{Tr}[(I_3)^3] = 0$, we have

$$\sum_f \mathrm{Tr}\left[I_{3f} Q_f^2\right] = \sum_f N_f \mathrm{Tr}\left[I_{3f}^2 Q_f\right] = \frac{1}{4}\sum_f N_f \mathrm{Tr}[Q_f] \tag{3.6}$$

One sees that Eq. (3.6) for the lepton or the quark multiplets alone do not vanish. But if all members of a family are included, we have

$$\text{Anomaly} \;\propto\; \sum_f N_f \mathrm{Tr}[Q_f] = 3 \times \left(Q_u + Q_d\right) + (Q_\nu + Q_e) = 0 \tag{3.7}$$

which satisfies the anomaly-free condition. It is hard to consider it an accident. If the correct theory is protected by the anomaly-free condition, the quark and the lepton have to be considered as members of one and the same group. In other words, the family structure is not accidental but essential for the construction of the right theory.

Incidentally, if one considers the inclusion of both chiralities in the anomaly calculation, the condition becomes

$$\mathrm{Tr}\left[t_a\{t_b, t_c\}\right]_L - \mathrm{Tr}\left[t_a\{t_b, t_c\}\right]_R = 0 \tag{3.8}$$

In the SM, the second term does not contribute. But in considering GUT groups, it needs not to be ignored. In fact, a model in which the second term is identical to the first (referred to as *vector type*) automatically satisfies the anomaly-free condition. For instance, $SO(10)$, which will be described later, has the left–right symmetry of this nature and is one of the attractive models.

3.3
SU(5)

If the grand unification is possible, all the coupling constants must agree with each other at some energy scale. The scale-dependent coupling constants can be calculated using the renormalization group equation (RGE) (see Section 7.2 of [2]) and, to the first loop correction, are expressed as

$$\begin{gathered}\frac{1}{\alpha_j(\mu)} = \frac{1}{\alpha_j(m_W)} + \frac{\beta_j}{2\pi}\ln\left(\frac{\mu}{m_W}\right) + \cdots \\ \begin{cases} \beta_1 = -\frac{1}{3}\left(4n_g + \frac{3}{10}\right) \\ \beta_2 = \;\;\frac{1}{3}\left(22 - 4n_g - \frac{1}{2}\right) \\ \beta_3 = \;\;\frac{1}{3}\left(33 - 4n_g\right) \end{cases}\end{gathered} \tag{3.9}$$

where $\alpha_j = g_j^2/(4\pi)$ are the coupling constants of the group $SU(j)$. μ is the energy scale at which their value is to be evaluated. Coefficients $\beta_j = \mu\partial\alpha_j/\partial\mu$ (referred to as *the β function*) can be obtained perturbatively by calculating loop diagrams that contribute to the coupling of the gauge boson. n_g is the number of generations whose members have mass below μ.

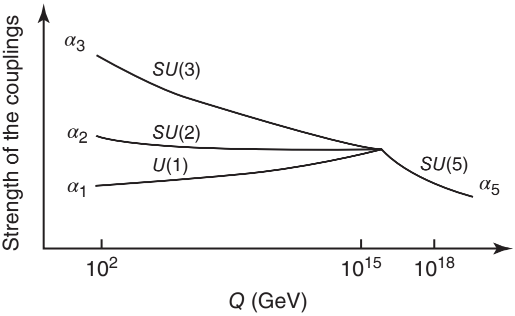


Figure 3.3 Evolution of coupling constants and the GUT scale. In order for the GUT (grand unified theory) to hold, the three coupling constants have to meet at some scale (together or separately).

In the non-Abelian groups, $\beta_j > 0$ and the asymptotic freedom holds. Substituting the measured values of $\alpha_j(m_W)$ in Eq. (3.9), evolution of each coupling constant is qualitatively described in Figure 3.3. The fact that all the three couplings approximately meet at a scale $\mu \sim 10^{15-16}$ prompted proposals of the GUTs [367, 368].

The energy scale is still below the Planck scale (~10^{19} GeV), and therefore a possibility of extending the unification to include the gravity remains. Many choices are available for the GUT group G, but here we discuss the simplest $SU(5)$ that satisfies the minimum requirement and is best suited to elucidate the basic concepts of the GUTs [368]. To make the theory consistent with experiments, a few modifications are necessary, which will be discussed later.

3.3.1
Fermion Representation

The group $SU(5)$ has $5^2 - 1 = 24$ generators and its rank is $5 - 1 = 4$ (see Appendix $SU(N)$ in [2]). Therefore, four generators can be diagonalized simultaneously. We choose λ_3, λ_8 of $SU(3)_{\text{color}}$, I_3 of $SU(2)$, and hypercharge Y of $U(1)$ in the SM and assign them to those diagonal generators. In order to find representations to which the fermions (quarks and leptons) are to be assigned, let us organize their quantum numbers and express them in terms of dimensions D of representations in $SU(3)$, $SU(2)$, and hypercharge of $U(1)$ like $(D_{SU(3)}, D_{SU(2)}, Y)$. The 15 fermions in the first generation are expressed in the Table 3.2.

The reason why we use charge conjugates d_L^c, u_L^c, e_L^c to express singlets d_R, u_R, e_R is to unify the handedness of all the fermions. This is necessary to assign all the fundamental fermions including the right-handed singlets to the same multiplet. For the moment, we pretend that the right-handed neutrino

$$\nu_L^c \quad (1, 1, 0) \tag{3.10}$$

does not exist to simplify arguments. Let us see whether the 15 fundamental fermions can be assigned to **5** and **10**, which are representations of minimum dimension in $SU(5)$ (see Table 3.4). It would be neat if we can assign all of them

Table 3.2 Quantum number assignment of the fundamental fermions.

Fermion	($SU(3)$, $SU(2)$, Y)	Multiplicity
(u_L, d_L)	$(3, 2, 1/3)$	6
d_L^c	$(3^*, 1, 2/3)$	3
u_L^c	$(3^*, 1, -4/3)$	3
(ν_L, e_L)	$(1, 2, -1)$	2
e_L^c	$(1, 1, 2)$	1

Note $f_L^c \equiv (f_R)^c$.

to a single multiplet **15**, but there is no reason not to do it either. Referring to Table 3.2, there are two possibilities to make $\mathbf{5}^*$. Using the index i ($i = 1-3$) for the color degrees of freedom, we have

$$\begin{aligned} \mathbf{5}^* &= (d_L^{ci}, \nu_L, e_L) = (3^*, 1, 2/3) \oplus (1, 2, -1) && (3.11a) \\ \text{or} \quad &= (u_L^{ci}, \nu_L, e_L) = (3^*, 1, -4/3) \oplus (1, 2, -1) && (3.11b) \end{aligned}$$

Which one to choose can be determined by considering the electric charge of the multiplet.

$$Q = I_3 + \frac{Y}{2} \tag{3.12}$$

If we assign two of the SU(5) generators to I_3 and Y (see Eq. (3.17) and Eq. (3.19)), both of them are traceless and hence, the electric charge Q should be traceless, too. This is in sharp contrast to the SM, where no constraint on Y (hence Q) could be imposed. The constraint of tracelessness means

$$\begin{aligned} &\sum_{a=1}^{5} Q_a = 3 \cdot Q(q^c) + Q_\nu + Q(e^-) = 0 \\ &\therefore \quad Q(q) = -Q(q^c) = -\frac{1}{3} \end{aligned} \tag{3.13}$$

Therefore, we have to use Eq. (3.11a). Thus the three q_L^c's that belong to $\mathbf{5}^*$ must be d_L^c. Writing the fields that belong to the fundamental representation as ψ_5^a ($a = 1 \sim 5$)[4] and taking the charge conjugate of Eq. (3.11a), the fundamental representation can be written as

$$\psi_5^a = \begin{bmatrix} d_R^1 \\ d_R^2 \\ d_R^3 \\ (e_L)^c \\ -(\nu_L)^c \end{bmatrix} = \left(3, 1, -\frac{2}{3}\right) \oplus (1, 2, 1) \tag{3.14a}$$

$$\overline{\psi}_{5a} \equiv \psi_5^{a\dagger}\gamma^0 = (\,\overline{d_R^1}, \overline{d_R^2}, \overline{d_R^3}, \overline{(e_L)^c}, -\overline{(\nu_L)^c}\,) \tag{3.14b}$$

---

[4] In the following, we write indices of $SU(5)$ group conveniently in the form of tensor formalism. Column vectors are written like contravariant vector, and row vectors like covariant ones. We adopt the convention to make sum when the same indices appear as both upper and lower ones.

The indices $a = 1 - 3$ correspond to the color R, G, B, and $a = 4, 5$ correspond to $I_3 = +1/2$ and $-1/2$, respectively. Thus we have defined the quantum numbers of each element in the fundamental representation **5**. The advantage of the GUT is already evident. The electric charge is quantized, and the reason why the electric charge of the d quark is one-third of that of the electron is clarified.

Quantum numbers of the larger dimensional representations can be determined by constructing direct products of the fundamental representation. For instance, $\mathbf{5} \otimes \mathbf{5} = \mathbf{15}_s \oplus \mathbf{10}_A$ (see Table 3.3), where suffixes S and A are attached to make the symmetric or antisymmetric nature of the wave functions explicit. Therefore, the representation **10** should behave like

$$\psi_{10}^{ab} \sim \psi_5^a \psi_5^b - \psi_5^b \psi_5^a \tag{3.15}$$

which is antisymmetric with respect to indices a, b.

Therefore, the part with $a, b = 1 - 3$ corresponds to $\mathbf{3}^*$ of the product representations $SU(3)$ $\mathbf{3} \otimes \mathbf{3} = \mathbf{6}_s \oplus \mathbf{3}_A^*$ which are u_L^{ci}'s. Similarly, the part with $a, b = 4, 5$ corresponds to e_L^c. Those with $a = 1 - 3, b = 4, 5$ belong to color **3** and isospin $I = 1/2$ which are (u_L, d_L). In summary, each member of ψ_{10}^{ab} of the 10-dimensional representation is expressed as

$$\psi_{10}^{ab} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & u^{c3} & -u^{c2} & -u^1 & -d^1 \\ -u^{c3} & 0 & u^{c1} & -u^2 & -d^2 \\ u^{c2} & -u^{c1} & 0 & -u^3 & -d^3 \\ u^1 & u^2 & u^3 & 0 & -e^c \\ d^1 & d^2 & d^3 & e^c & 0 \end{bmatrix}_L \tag{3.16}$$

where $1/\sqrt{2}$ is a normalization constant and the index L denotes that the fermions in the representation are left-handed. Thus, we were able to assign all the 15 fundamental fermions carrying the correct quantum numbers of the SM to **5** and **10** with no redundancy. As there is no space left for the right-handed $(\nu_L)^c$, the only way to include it is to assign it to a singlet. In Table 3.4, we list the basic $SU(5)$ representations and their quantum numbers in the decomposed form of $SU(3) \times SU(2) \times U(1)$.

Table 3.3 Decomposition[a] of product representations in $SU(5)$.

$5 \otimes 5^*$	$= 1 \oplus 24$
$5 \otimes 5$	$= 10_A \oplus 15_S$
$5^* \otimes 10$	$= 5 \oplus 45$
$5 \otimes 10$	$= 40 \oplus 10^*$
$10 \otimes 10$	$= 5^* \oplus 45^* \oplus 50$
$10^* \otimes 10$	$= 1 \oplus 24 \oplus 75$

[a] See Appendix G of [1] for derivation.

Table 3.4 Representation of $SU(5)$ in terms of $SU(3) \times SU(2) \times U(1)$.

$SU(5)$ representation	$(SU(3), SU(2), Y)$	Comment
5	$(3, 1, -2/3) \oplus (1, 2, 1)$	Fundamental
10	$(3, 2, 1/3) \oplus (3^*, 1, -4/3) \oplus (1, 1, 2)$	Antisymmetric
15	$(6, 1, -4/3) \oplus (3, 2, 1/3) \oplus (1, 3, 2)$	Symmetric
$24 = 24^*$	$(8, 1, 0) \oplus (3, 2, -5/3) \oplus (3^*, 2, 5/3) \oplus (1, 3, 0) \oplus (1, 1, 0)$	Adjoint
45	$(8, 2, 1) \oplus (6^*, 1, -2/3) \oplus (3, 3, -2/3) \oplus (3^*, 2, -7/3)$ $\oplus (3, 1, -2/3) \oplus (3^*, 1, 8/3) \oplus (1, 2, 1)$	
50	$(8, 2, 1) \oplus (6, 1, 8/3) \oplus (6^*, 3, -2/3) \oplus (3^*, 2, -7/3)$ $\oplus (3, 1, -2/3) \oplus (1, 1, -4)$	

3.3.2
Representation of the Gauge Particle

Adjoint representation: Next, let us construct generators $[L_i]^a_b$ ($i = 1 - 24$, a, $b = 1 - 5$) of the $SU(5)$ group. According to the fundamental representation Eq. (3.14), indices $a, b = 1 - 3$ correspond to colors and $a, b = 4, 5$ to the isospin. Therefore, we can assign $i = 1 - 8$ to QCD generators, and $i = 9 - 11$ to those of $SU(2)$.

$$L_i\,(i = 1 \sim 8) = \left[\begin{array}{ccccc} & & & 0 & 0 \\ & \lambda_i & & 0 & 0 \\ & & & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right], \quad L_i\,(i = 9 \sim 11) = \left[\begin{array}{ccccc} & & & 0 & 0 \\ & \mathbf{0} & & 0 & 0 \\ & & & 0 & 0 \\ 0 & 0 & 0 & & \\ 0 & 0 & 0 & \multicolumn{2}{c}{\sigma_j} \end{array}\right] \tag{3.17}$$

where λ_i are Gell-Mann matrices and $j = 1 - 3$ correspond to $i = 9 - 11$. These generators satisfy the standard normalization condition

$$\mathrm{Tr}\left[L_i L_j\right] = 2\delta_{ij} \tag{3.18}$$

As we have three diagonal generators L_3, L_8, $L_{11}\,(= 2I_3)$, we choose to assign the fourth diagonal generator to $L_{12} = cY$. Referring to the hypercharge of ψ_5 and using the normalization condition Eq. (3.18), we can fix the constant c.

$$L_{12} = \sqrt{\frac{3}{5}}\,Y = \sqrt{\frac{3}{5}} \begin{bmatrix} -2/3 & 0 & 0 & 0 & 0 \\ 0 & -2/3 & 0 & 0 & 0 \\ 0 & 0 & -2/3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.19}$$

The remaining 12 generators $L_{13} - L_{24}$ can be constructed in the following way.

$$L_{13} = \begin{bmatrix} & & & 1 & 0 \\ & \mathbf{0} & & 0 & 0 \\ & & & 0 & 0 \\ 1 & 0 & 0 & & \\ & & & \multicolumn{2}{c}{0} \\ 0 & 0 & 0 & & \end{bmatrix} \qquad L_{14} = \begin{bmatrix} & & & i & 0 \\ & \mathbf{0} & & 0 & 0 \\ & & & 0 & 0 \\ -i & 0 & 0 & & \\ & & & \multicolumn{2}{c}{0} \\ 0 & 0 & 0 & & \end{bmatrix} \tag{3.20a}$$

$$L_{15} = \begin{bmatrix} & & & 0 & 1 \\ & \mathbf{0} & & 0 & 0 \\ & & & 0 & 0 \\ 0 & 0 & 0 & & \\ & & & \multicolumn{2}{c}{0} \\ 1 & 0 & 0 & & \end{bmatrix} \qquad L_{16} = \begin{bmatrix} & & & 0 & -i \\ & \mathbf{0} & & 0 & 0 \\ & & & 0 & 0 \\ 0 & 0 & 0 & & \\ & & & \multicolumn{2}{c}{0} \\ i & 0 & 0 & & \end{bmatrix} \tag{3.20b}$$

$L_{17} - L_{20}$ and $L_{21} - L_{24}$ can be constructed in a similar manner. Note that $L_{13} - L_{24}$ are operators to exchange quarks and leptons. They do not exist in the SM but have appeared in the GUTs.

The gauge boson V_μ belongs to the adjoint representation **24** and can be expressed as follows:

$$\frac{1}{\sqrt{2}} V_\mu = \sum_{i=1}^{24} V_\mu^i \frac{L_i}{2} \tag{3.21a}$$

$$V_\mu = \begin{bmatrix} G_1^1 - \dfrac{2B}{\sqrt{30}} & G_2^1 & G_3^1 & \overline{X}_1 & \overline{Y}_1 \\ G_1^2 & G_2^2 - \dfrac{2B}{\sqrt{30}} & G_3^2 & \overline{X}_2 & \overline{Y}_2 \\ G_1^3 & G_2^3 & G_3^3 - \dfrac{2B}{\sqrt{30}} & \overline{X}_3 & \overline{Y}_3 \\ X^1 & X^2 & X^3 & \dfrac{W^0}{\sqrt{2}} + \dfrac{3B}{\sqrt{30}} & W^+ \\ Y^1 & Y^2 & Y^3 & W^- & -\dfrac{W^0}{\sqrt{2}} + \dfrac{3B}{\sqrt{30}} \end{bmatrix} \tag{3.21b}$$

where

$$\begin{aligned} G_2^1 &= \frac{(g_1 - ig_2)}{\sqrt{2}}, \quad G_1^2 = \frac{(g_1 + ig_2)}{\sqrt{2}}, \\ G_3^1 &= \frac{(g_4 - ig_5)}{\sqrt{2}}, \quad G_1^3 = \frac{(g_4 + ig_5)}{\sqrt{2}}, \\ G_3^2 &= \frac{(g_6 - ig_7)}{\sqrt{2}}, \quad G_2^3 = \frac{(g_6 + ig_7)}{\sqrt{2}} \\ G_1^1 &= \frac{g_3}{\sqrt{2}} + \frac{g_8}{\sqrt{6}}, \quad G_2^2 = -\frac{g_3}{\sqrt{2}} + \frac{g_8}{\sqrt{6}}, \quad G_3^3 = -\sqrt{\frac{2}{3}} g_8 \end{aligned} \tag{3.21c}$$

g_i $(i = 1 - 8)$ represents the eight gluons. X, Y are the new gauge bosons that have appeared in the GUTs. They convert quarks to leptons and vice versa. In order to see their quantum numbers more explicitly, we use $\mathbf{24} \sim \mathbf{5} \otimes \mathbf{5}^*$ and we express them in terms of equivalent fermions that have the same effective role.

$$X \sim (e_L)^c d_L^c \text{ or } u_R u_L, \qquad Y \sim (\nu_L)^c d_L^c \text{ or } u_R d_L \tag{3.22}$$

That is, the gauge bosons (X, Y) constitute an isospin doublet, and have the color quantum number as well as the lepton number. Because of this, they are sometimes referred to as *leptoquarks*. In terms of $(SU(3), SU(2))$ dimensionality,

$$X, Y = (3^*, 2) \tag{3.23}$$

and in terms of the electric charge Q, isospin I_3, and the hypercharge Y

$$\begin{aligned} X(Q, I_3, Y) &= \left(\frac{4}{3}, \ \frac{1}{2}, \frac{5}{3}\right) \\ Y(Q, I_3, Y) &= \left(\frac{1}{3}, -\frac{1}{2}, \frac{5}{3}\right) \end{aligned} \tag{3.24}$$

Knowing the gauge field, the Lagrangian of interacting fermion fields can be written down.

$$\mathscr{L} = \mathscr{L}_{\text{gauge}} + \mathscr{L}_5 + \mathscr{L}_{10} \tag{3.25a}$$

$$= -\frac{1}{4}\mathrm{Tr}\left[f^{\mu\nu} f_{\mu\nu}\right] + \overline{\psi}_5 i\gamma^\mu D_\mu \psi_5 + \overline{\psi}_{10\,ab} i\gamma^\mu (D_\mu \psi_{10})^{ab} \tag{3.25b}$$

$$\begin{aligned} &= -\frac{1}{4}(f^{\mu\nu})^b_a (f_{\mu\nu})^a_b + \overline{\psi}_{5\,a} i\gamma^\mu \left\{ \delta^a_b \partial_\mu + i\frac{g_5}{\sqrt{2}}(V_\mu)^a_b \right\} \psi_5^b \\ &\quad + \overline{\psi}_{10\,ab} i\gamma^\mu \left\{ \delta^a_c \partial_\mu + i\frac{2g_5}{\sqrt{2}}(V_\mu)^a_c \right\} \psi_{10}^{cb} \end{aligned} \tag{3.25c}$$

$$(f_{\mu\nu})^a_b = \partial_\mu (V_\nu)^a_b + \frac{ig_5}{\sqrt{2}}(V_\mu)^a_c (V_\nu)^c_b - (\mu \leftrightarrow \nu) \tag{3.25d}$$

where g_5 is the coupling constant of the gauged $SU(5)$.

Proof of covariant derivative:

Obtaining the covariant derivative acting on ψ_5 is straightforward; that acting on ψ_{10} can be proved as follows: We denote representations of $\mathbf{5}$ as ψ^a, of $\mathbf{10}$ as χ^{ab}, and of the generators in the ψ space as $L_{A\,b}^{\ a}$, and those in the χ space as $L_{A\,cd}^{\ ab}$ where $A = 1 \sim 24$, $a, b = 1 - 5$. As χ^{ab} transforms like $(\psi^a \psi^b - \psi^b \psi^a)$, the infinitesimal transformation can be expressed as

$$\psi + \delta\psi = \left(1 - \frac{ig_5}{2}\varepsilon^A L_A\right)\psi = \left(\delta^a_b - \frac{ig_5}{2}\varepsilon^A L_{A\,b}^{\ a}\right)\psi^b \tag{3.26a}$$

$$\begin{aligned} \chi^{ab} + \delta\chi^{ab} &= \left(\delta^{ab}_{cd} - \frac{ig_5}{2}\varepsilon^A L_{A\,cd}^{\ ab}\right)\chi^{cd} \\ &= \frac{1}{2}\left(\delta^a_c - \frac{ig_5}{2}\varepsilon^A L_{A\,c}^{\ a}\right)\left(\delta^b_d - \frac{ig_5}{2}\varepsilon^A L_{A\,d}^{\ b}\right)\chi^{cd} \end{aligned} \tag{3.26b}$$

From the above-mentioned equations, we derive

$$L_{A}{}^{ab}_{cd} = \frac{1}{2}\left(\delta^a_c L_{A}{}^{b}_{d} + \delta^b_d L_{A}{}^{a}_{c} - \delta^a_d L_{A}{}^{b}_{c} - \delta^b_c L_{A}{}^{a}_{d}\right) \tag{3.27}$$

In the Lagrangian, it appears in a bilinear form, that is,

$$\overline{\chi}_{ab} L_{A}{}^{ab}_{cd} \chi^{cd} = 2\overline{\chi}_{ac} L_{A}{}^{c}_{b} \chi^{ab} = 2\overline{\chi}_{ab} L_{A}{}^{a}_{c} \chi^{cb} \tag{3.28}$$

Using matrix representation for the gauge field $V_\mu = V^A_\mu L_A/\sqrt{2}$, the covariant derivative for χ^{bc} can be written as

$$\mathscr{L}_{10} = \overline{\chi}_{ab} i\gamma^\mu \left[\delta^a_c \partial_\mu + \frac{2ig_5}{\sqrt{2}}(V_\mu)^a_c\right] \chi^{cb} \tag{3.29}$$

■

Weinberg angle: From Eq. (3.25c), one can immediately derive that

$$g_{\text{color}}\,(\equiv g_3) = g_W\,(= g_2) = g_5 \tag{3.30}$$

Notice that the above relations hold at the GUT unification energy scale, namely at $\mu \simeq 10^{15-16}$ GeV ($\equiv M_{\text{GUT}}$), which will be referred to as the *GUT scale*. Next, in order to determine the coupling constant of the $U(1)$ gauge boson B, we extract terms that contain neutral gauge bosons W^0_μ and B_μ and rewrite them in terms of the photon field A_μ and Z_μ:

$$\begin{aligned}\frac{g_5}{2}\left[W^0 L_{11} + BL_{12}\right] &= \frac{g_5}{2}\left[\left(\hat{s}_W A + \hat{c}_W Z\right) L_{11} + \left(\hat{c}_W A - \hat{s}_W Z\right) L_{12}\right] \\ &= \frac{g_5}{2}\left[\left(\hat{s}_W L_{11} + \hat{c}_W L_{12}\right) A + \left(\hat{c}_W L_{11} - \hat{s}_W L_{12}\right) Z\right] \\ &\equiv eQA + g_Z Q_Z Z\end{aligned} \tag{3.31}$$

where $\hat{s}_W = \sin\hat{\theta}_W = \sin\hat{\theta}_{W\overline{MS}}$, $\hat{c}_W = \cos\hat{\theta}_W$ are the Weinberg angles defined in the modified minimum subtraction scheme (see Section 5.6 of [2]). From the equation, we can express the electric charge operator as

$$eQ = \frac{g_5}{2}\left(\hat{s}_W L_{11} + \hat{c}_W L_{12}\right) \tag{3.32a}$$

On the other hand, it can also be expressed as $Q = I_3 + Y/2$. Therefore, using $I_3 = L_{11}/2$, $Y = \sqrt{5/3}L_{12}$, we have

$$Q = \frac{1}{2}\left(L_{11} + \sqrt{\frac{5}{3}}L_{12}\right) \tag{3.32b}$$

Comparing Eq. (3.32a) and (3.32b), we finally obtain

$$e = g_5 \sin\hat{\theta}_W \tag{3.33a}$$

$$g_B = \frac{e}{\cos\hat{\theta}_W} = \sqrt{\frac{3}{5}}g_5 = g_5 \tan\hat{\theta}_W \tag{3.33b}$$

$$\sin\hat{\theta}_W = \sqrt{\frac{3}{8}} \tag{3.33c}$$

Thus the three coupling constants in the SM are connected by the relation

$$g_{\text{color}}\,(=g_3)=g_W\,(=g_2)=\sqrt{\frac{5}{3}}g_B\,(=g_1)=g_5 \tag{3.34}$$

We remind the reader again that these relations are true only at the GUT scale.

3.3.3
Symmetry Breakdown

The X, Y gauge bosons induce transitions between the quarks and leptons, which means that the proton can decay. However, we know that the proton is stable. In order not to contradict the observations, the decay life of the proton must be long. From dimensional arguments, the decay rate is expressed as $\Gamma \sim \frac{g_5^4}{M_X^4}(m_p)^5$; that is, the masses M_X, M_Y of the X, Y must be very large. Therefore, we assume that the $SU(5)$ symmetry is spontaneously broken in two stages, one at the GUT scale ($\mu \simeq M_{\text{GUT}} \simeq 10^{15}$ GeV) when X, Y obtain their mass, and the second at the electroweak scale ($\mu \simeq 250$ GeV) when W, Z obtain their masses.

To induce spontaneous symmetry breakdown, we need to add the Higgs field to the Lagrangian. The first stage of breaking can be achieved through real scalars that belong to adjoint representation, i.e., **24**. As $\mathbf{24} \oplus \mathbf{1} = \mathbf{5} \otimes \mathbf{5}^*$, Φ can be represented by a 5×5 matrix similar to the gauge boson expression Eq. (3.21b). In order to give mass to X, Y and not to others, one can choose the Higgs Φ^a_b in **24** to have a vacuum expectation value (VEV) in a form [369, 370].

$$<\Phi>=\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & -3/2 & \\ & & & & -3/2 \end{bmatrix} V \tag{3.35}$$

To see whether it behaves as desired, we calculate the covariant derivative first.

$$\Phi_i\,(i=1\sim 24)=\Phi^a_b\,(a,b=1\sim 5)\sim \psi_5^a\psi^*_{5\,b}-\frac{1}{5}\delta^a_b\psi_5^c\psi_{5c}{}^* \tag{3.36}$$

In this form, the transformation property of Φ can be derived from that of ψ_5, ψ_5^*. Using

$$\psi_5^a \to \psi_5^a+\delta\psi_5^a=\psi_5^a-i\frac{g_5}{2}\varepsilon^A L_{Ac}{}^a\psi_5^c \tag{3.37a}$$

$$\psi^*_{5\,b}\to\psi^*_{5\,b}+\delta\psi^*_{5\,b}=\psi^*_{5\,b}+i\frac{g_5}{2}\varepsilon^A\psi^*_{5\,c}L_{Ab}{}^c \tag{3.37b}$$

we derive the infinitesimal transformation of Φ^a_b as

$$\Phi^a_b\to\Phi^a_b-i\frac{g_5}{2}[L_A,\,\Phi]^a_b \tag{3.38}$$

and its covariant derivative

$$(D_\mu\Phi)^a_b=\partial_\mu\Phi^a_b+i\frac{g_5}{\sqrt{2}}[V_\mu,\,\Phi]^a_b \tag{3.39}$$

The kinetic energy of the Higgs can be expressed as

$$\mathscr{L}_{\Phi,KE} = \frac{1}{2}\mathrm{Tr}\left[(D_\mu\Phi)^\dagger(D^\mu\Phi)\right] \tag{3.40}$$

By replacing Φ with its VEV, the mass term of the gauge boson can be obtained.

$$\begin{aligned}\mathscr{L}_{\mathrm{mass}} &= \frac{1}{2}\mathrm{Tr}\left[(D_\mu\Phi)^\dagger(D^\mu\Phi)\right]_{\Phi=<\Phi>} \\ &= \frac{g_5^2}{4}\mathrm{Tr}\left([V_\mu, <\Phi>]^\dagger[V^\mu, <\Phi>]\right) \equiv -\frac{1}{2}m_{ij}^2 V_\mu^i V^{\mu j*}\end{aligned} \tag{3.41}$$

If $<\Phi>$ and V_μ are of the form

$$<\Phi> \sim \begin{bmatrix} p & & & & \\ & p & & \mathbf{0} & \\ & & p & & \\ & \mathbf{0} & & q & \\ & & & & q \end{bmatrix}, \qquad V_\mu \sim \left[\begin{array}{c|c} 3\times 3 & A \\ \hline B & 2\times 2 \end{array}\right] \tag{3.42}$$

we have

$$[V_\mu, <\Phi>] \sim \left[\begin{array}{c|c} 0 & -(p-q)A \\ \hline (p-q)B & 0 \end{array}\right] \tag{3.43}$$

Therefore only X, Ys in the V_μ^i of Eq. (3.41) get the VEV as desired. Setting $p = V$, $q = -3V/2$, the masses of X, Y are given by

$$M_X^2 = M_Y^2 = \frac{25g_5^2}{8}V^2 \simeq 0.94V^2 \tag{3.44}$$

As to the value of g_5, we refer to Eq. (3.53) later. As the value of the V is of the order of the GUT scale, we have $M_X \simeq 10^{15}$ GeV. The potential part of the Higgs Lagrangian can be constructed similarly to that of the SM, which has the form

$$\mathscr{L}_{\mathrm{SM,\,PE}}(|\phi|^2) = \lambda\left(|\phi|^2 - \frac{\mu^2}{2\lambda}\right)^2 \tag{3.45}$$

In order to break symmetry spontaneously, the potential has to include a quartic term and in general has a form

$$\mathscr{L}_{\mathrm{PE}}(|\Phi|^2) = -\frac{\mu^2}{2}\mathrm{Tr}[\Phi^2] + \frac{a}{4}\left(\mathrm{Tr}[\Phi^2]\right)^2 + \frac{b}{2}\mathrm{Tr}[\Phi^4] \tag{3.46}$$

A term $\sim\Phi^3$ is allowed but we omit it for simplicity. If $b > 0$, $\mu^2 > 0$, and $a > 7b/5$, the minimum can be defined uniquely to give

$$V^2 = \frac{2\mu^2}{15a + 7b} \tag{3.47}$$

Our next question is how to have VEV to give mass to W, Z. Is it possible to arrange Φ such that the electroweak symmetry breaking can also be induced by Φ?

The answer is no. One reason is that there exist no color singlet and isospin 1/2 components in the Φ, and hence it cannot give VEV to the W, Z (see Table 3.4). Another reason is that the fermion mass term is of the form $[\psi^c{}_L C\psi_L + h.c.]$. As the fermions belong to either $\mathbf{5}^*$ or **10**, according to Table 3.4, the mass term has to belong to one of

$$\begin{aligned}
&\mathbf{5}^* \otimes \mathbf{5}^* = \mathbf{10}^* \oplus \mathbf{15}^* \\
&\mathbf{5}^* \otimes \mathbf{10} = \mathbf{5} \oplus \mathbf{45} \\
&\mathbf{10} \otimes \mathbf{10} = \mathbf{5}^* \oplus \mathbf{45}^* \oplus \mathbf{50}
\end{aligned} \tag{3.48}$$

and cannot couple to **24**. Therefore, the fermions cannot acquire mass if only Φ has the VEV. This is the reason why two-step symmetry breaking is necessary. Referring to Eq. (3.48), the second Higgs can belong to either **5** or **45**. The simplest version of $SU(5)$ uses **5**. We denote this Higgs as H_5. As it contains the $I = 1/2$ component, we may choose VEV of H_5 as

$$< H_5 >= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \frac{v}{\sqrt{2}} \end{bmatrix} = (3, 1, Y) \oplus (1, 2, Y) \tag{3.49}$$

so that gauge bosons of the SM can acquire the necessary mass spectrum. The mass giving Yukawa interaction has the form

$$u \leftarrow [\mathbf{10}] \times [\mathbf{10}] \times [H_5], \qquad d \,\text{or}\, e \leftarrow [\mathbf{5}^*] \times [\mathbf{10}] \times [H_5^*] \tag{3.50}$$

An explicit form of the interaction will be given later in Eqs. (3.60) and (3.61).

In summary, it is possible to construct a model of GUTs that includes the SM as its part. Whether it describes Nature is another problem, which is our next discussion item.

3.3.4
Predictions

Weinberg Angle We derived a value of the Weinberg angle in Eqs. (3.30) and (3.33). However, it holds only at the GUT scale ($E \sim M_{\text{GUT}} \sim 10^{15}$ GeV) and has to be corrected to give right behavior at low energies. The correction can be obtained using the RGE as shown in Eq. (3.9). In order to see what kind of values the coupling constants take at low energies, we set up evolution equations for them starting from $\mu = M_{\text{GUT}}$. Substituting expressions for the β function in each group [361–363], we have

$$\frac{1}{\alpha_3(\mu)} = \frac{1}{\alpha_{\text{GUT}}} + \frac{1}{2\pi}\left(11 - \frac{4}{3}n_g\right)\ln\left(\frac{\mu}{M_{\text{GUT}}}\right) + \cdots \tag{3.51a}$$

$$\frac{1}{\alpha_2(\mu)} = \frac{\sin^2\hat{\theta}_W}{\alpha_{\text{EM}}} = \frac{1}{\alpha_{\text{GUT}}} + \frac{1}{2\pi}\left(\frac{22}{3} - \frac{4}{3}n_g - \frac{N_H}{6}\right)\ln\left(\frac{\mu}{M_{\text{GUT}}}\right) + \cdots \tag{3.51b}$$

$$\frac{1}{\alpha_1(\mu)} = \frac{3}{5}\frac{\cos^2\hat{\theta}_W}{\alpha_{\text{EM}}} = \frac{1}{\alpha_{\text{GUT}}} - \frac{1}{2\pi}\left(\frac{4}{3}n_g + \frac{N_H}{10}\right)\ln\left(\frac{\mu}{M_{\text{GUT}}}\right) + \cdots \tag{3.51c}$$

$$\alpha_{\text{GUT}} = \alpha_i(M_{\text{GUT}}) = \frac{g_5^2}{4\pi} \tag{3.51d}$$

where n_g is the number of generations whose members have $m_f < \mu$ and N_H is the number of Higgs doublets with $m_H < \mu$. Assuming $n_g = 3$, $N_H = 1$, one gets

$$\sin^2\hat{\theta}_W(m_Z) = \frac{3}{8} - \frac{109}{48\pi}\alpha_{\text{EM}}(m_Z)\ln\left(\frac{M_{\text{GUT}}}{m_Z}\right) \tag{3.52a}$$

$$\frac{\alpha_{\text{EM}}(m_Z)}{\alpha_3(m_Z)} = \frac{3}{8} - \frac{67}{16\pi}\alpha_{\text{EM}}(m_Z)\ln\left(\frac{M_{\text{GUT}}}{m_Z}\right) \tag{3.52b}$$

As we have the measured values of $\alpha_{\text{EM}}(m_Z)$, $\alpha_3(m_Z)$, we can determine M_{GUT} from Eq. (3.52b), which gives $\sim 10^{15}$ GeV. Substituting the value of M_{GUT} in Eq. (3.52a), we can calculate the value of $\sin^2\hat{\theta}_W(m_Z)$ based on the GUT $SU(5)$ model. The RGEs in Eq. (3.51) are based on one-loop calculation. By including higher order radiative corrections, one can obtain better values. Using a two-loop equation, one obtains [371–373].

$$\alpha_{\text{GUT}} = 0.0244 \pm 0.0002 \tag{3.53a}$$

$$\sin^2\hat{\theta}_W(m_Z) = 0.214 \pm 0.004 \tag{3.53b}$$

These values are close to but do not quite agree with the LEP precision data,

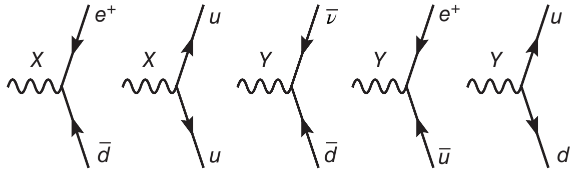
$$\sin^2\hat{\theta}_W = 0.23119 \pm 0.00014 \quad [7, 374] \tag{3.54}$$

Proton Decay Here we derive the decay rates of the proton in the $SU(5)$ GUTs. The gauge interaction of the fermions is given by Eq. (3.25). Using Eqs. (3.14) and (3.16), we can write down the interaction terms explicitly. Picking terms containing X, Y, we have

$$\begin{aligned}-\mathscr{L}_{XY} = \frac{g_5}{\sqrt{2}}\Big[&\overline{X}^i_\mu(\overline{d}_{iR}\gamma^\mu e^c{}_R + \overline{d}_{iL}\gamma^\mu e^c{}_L - \varepsilon_{ijk}\overline{u}^c{}_L^j\gamma^\mu u_L{}^k)\\ &-\overline{Y}^i_\mu(\overline{d}_{iR}\gamma^\mu \nu^c{}_R + \overline{u}_{iL}\gamma^\mu e^c{}_L - \varepsilon_{ijk}\overline{u}^c{}_L^j\gamma^\mu d_L{}^k) + (h.c.)\Big]\end{aligned} \tag{3.55}$$

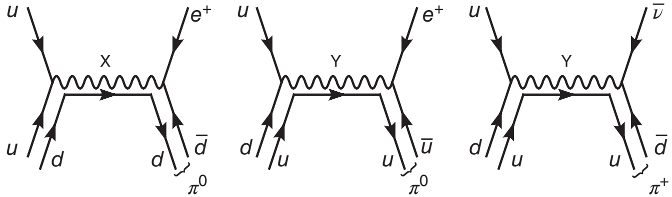
where $ijk = 1 - 3$ denote the color degrees of freedom. Feynman diagrams for the X, Y interactions are shown in Figure 3.4.

One notices that X and Y break the baryon number (B) as well as the lepton number (L) conservation law, but $B - L$ is conserved. The proton and the neutron can decay to leptons via processes depicted in Figure 3.5. The fact that diagrams



B	−1/3	2/3	−1/3	−1/3	2/3
L	−1	0	−1	−1	0
$B-L$	2/3	2/3	2/3	2/3	2/3
	(a)	(b)	(c)	(d)	(e)

Figure 3.4 $B-L$ is conserved in the gauge coupling of $X(Q=4/3)$ and $Y(Q=1/3)$ with fermions.



(a) (b) (c)

Figure 3.5 Feynman diagrams for the proton decay. (a) and (b) $p \to e^+\pi^0$. (c) $p \to \bar{\nu}\pi^+$.

of the process $p \to e^+ + \pi^0$ are easy to draw means that, if the proton actually decays, those diagrams are most likely the dominant processes. This is one of the conspicuous characteristics of the $SU(5)$ model. Many other models give similar predictions [361–363]:

$$\begin{array}{lll} p & \to & e^+ + \pi^0 \quad \sim 40-60\% \\ p & \to & e^+ + \omega \quad \sim 5-20\% \\ p & \to & e^+ + \rho^0 \quad \sim 1-10\% \\ p & \to & \bar{\nu}_e + \pi^+ \quad \sim 16-24\% \end{array} \tag{3.56}$$

In summary, dominance of the $p \to e^+\pi^0$ is a common feature to many models.

When the mass of the gauge bosons X, Y is large, the proton decay can be approximated by the four-Fermi interaction. Then the decay formula has the same form as that of muon decay. The lifetime can be expressed as

$$\tau_p = \tau_\mu \left(\frac{m_\mu}{m_p}\right)^5 \left(\frac{M_X}{m_W}\right)^4 \frac{BR(p \to e\pi^0)}{F} \sim \frac{M_X^4}{\alpha_5^2 m_p^5} \tag{3.57}$$

F is a model-dependent correction factor due to the quark wave function inside the proton. It has a large uncertainty, but is considered to be of order of $\sim O(1)$. Calculations show [361–363]

$$\tau_p(p \to e^+\pi^0) \equiv \frac{\tau_p}{BR(p \to e\pi)} \approx 4 \times 10^{29\pm0.7}\text{years} \tag{3.58}$$

Observations show

$$\begin{aligned}\frac{\tau_p}{BR(p \to e + \pi^0)} &> 8.2 \times 10^{33}\text{years} \qquad [375]\\ \frac{\tau_n}{BR(n \to e^+\pi^-)} &> 1.6 \times 10^{32}\text{ years} \qquad [376]\end{aligned} \tag{3.59}$$

which rule out the minimal $SU(5)$ model we have discussed so far. In fact, the experimental results rule out most non-SUSY GUT models [377], as the decay $N \to e\pi$ is a prominent feature of such models.

Scale dependence of the quark mass In the GUTs, quarks and leptons are in the same multiplets, and hence mass relations among them can be derived. In a special case where the mass is generated by the mechanism expressed in Eq. (3.50), the mass Lagrangian is written down as

$$\mathscr{L}_{\text{YUKAWA}} = g_u \varepsilon_{ijklm} \overline{\Psi^{cij}_{10\,L}} \Psi^{kl}_{10\,L} H^m_5 + g_d \overline{\Psi_{5R\,a}} \Psi^{ab}_{10\,L} H^\dagger_{5\,b} + (h.c.) \tag{3.60}$$

Substitute the VEV of H_5 as given by Eqs. (3.49) into (3.60) and we obtain:

$$\begin{aligned}\mathscr{L}_d &= \frac{g_d v}{\sqrt{2}} \left[\overline{\Psi}_{5\,a} \Psi^{a5}_{10} + h.c.\right] = -\frac{g_d v}{\sqrt{2}} \left[\overline{d^i_R} d_{L\,i} + \overline{e^c_R} e^c{}_L + h.c.\right]\\ &= -\frac{g_d v}{\sqrt{2}} \left[\sum \overline{d}d + \overline{e}e\right]\end{aligned} \tag{3.61}$$

We conclude that $m_e = m_d$. That is, in the minimal $SU(5)$ model, we have the relations

$$m_d = m_e, \quad m_s = m_\mu, \quad m_b = m_\tau \tag{3.62}$$

These relations are valid at the GUT scale. As the mass generated by the Higgs mechanism is proportional to the Higgs–fermion coupling strength, it runs as a function of the energy scale μ. We can again use the RGE to calculate their strength at low energies to obtain [370] [see also Eq. (7.101) of [2]]

$$\frac{m_b(\mu_1)}{m_\tau(\mu_1)} = \frac{m_b(\mu_2)}{m_\tau(\mu_2)} \left[\frac{\alpha_3(\mu_1)}{\alpha_3(\mu_2)}\right]^{4/[11-(4n_g/3)]} \left[\frac{\alpha_1(\mu_1)}{\alpha_1(\mu_2)}\right]^{4n_g/3} \tag{3.63}$$

Taking $\mu_1 \sim 10\,\text{GeV}$, $\mu_2 \sim M_X$, one gets a value $\simeq 2.5$, which agrees with experimental observation for m_b/m_τ. Values of m_s/m_μ, m_d/m_e do not agree with observation, but it is possible to ascribe the difference to nonperturbative effects because of the smallness of the mass. Such calculations alone do not validate the models, but are useful in constructing them.

In summary, out of three parameters we considered, two give inconsistent values with experiments. We conclude that the $SU(5)$ GUT is not enough to be a viable model at least in its minimal version. If the GUT is conceptually correct, we need further improvements.

3.4
SO(10)

3.4.1
Left–Right Symmetric World

Before we rectify the deficits of $SU(5)$ GUT and proceed to discuss the SUSY, we mention the possibilities of other groups which are generically referred to as *non-SUSY GUTs*. They are chosen to make GUT models more attractive phenomenologically, but the basic features we discussed in the previous section are inherent to most of the non-SUSY GUTs.

One old but intriguing question is whether the left–right disparity in the SM is an inherent characteristic of Nature. It could be superficial and a possibility exists that the left–right symmetry is recovered at high energies. If we adopt this view, it could be recovered at the GUT scale. The simplest way to recover it is to extend the $SU(2)_L$ in the SM to $SU(2)_L \times SU(2)_R$ and require existence of the gauge boson W_R which couples to the right-handed particles (see Section 2.1.2). If the mass of W_R is much larger than that of the standard W, its existence does not contradict the observations. Among the groups that contain $SU(2)_L \times SU(2)_R$ as a subgroup, $SO(10)$ and E_6 are the most popular. $SO(10)$ is a group of elements to keep the length of 10-dimensional real space invariant. E_6 is referred to as the *exceptional group* and there is no regularity like that of $SU(N)$ or $SO(N)$. One reason why E_6 is popular is that it appears as a low-energy limit of superstring theories. The two groups are related and can be decomposed in various ways to

$$\begin{aligned} E_6 \supset SO(10) &\to SU(5) \times U(1) \\ &\to SU(4) \times SU(2)_L \times SU(2)_R \\ &\to SU(3) \times SU(2)_L \times SU(2)_R \times U(1) \\ &\to SU(4) \times SU(2)_L \times U(1) \end{aligned} \tag{3.64}$$

If the breaking takes the first path in Eq. (3.64) through $SU(5)$, what we have discussed so far applies equally to the $SO(10)$ model. If Nature has chosen the second or the third path via $SU(2)_L \otimes SU(2)_R$ intermediate state, the left–right symmetry, that is, parity recovery, is natural.

The second attractive feature of the $SO(10)$ is that it contains a 16-dimensional spinor representation. All the 16 fermions can be contained in a single representation with the right quantum number, which is an aesthetically attractive feature. One can express this representation in two ways: $SU(5) \times U(1)$, or $SU(2)_L \times SU(2)_R \times SU(4)_c$.

$$\mathbf{16}\,[SO(10)] = \mathbf{10} \oplus \mathbf{5}^* \oplus \mathbf{1} \qquad [SU(5) \times U(1)] \tag{3.65a}$$

$$= (2, 1, 4) \oplus (1, 2, 4^*) \qquad SU(2)_R \times SU(2)_L \times SU(4)_c \tag{3.65b}$$

Here, in $SU(4)_c$ there are four degrees of freedom, of which the first three are assigned to the color degrees of freedom and the fourth to the lepton degree of freedom. Table 3.5 shows expressions of Eq. (3.65) explicitly in terms of fermions.

Table 3.5 Sixteen fermions are grouped in terms of $SU(5) \times U(1)$ multiplets ($\mathbf{5}^* \oplus \mathbf{10} \oplus \mathbf{1}$: left table) as well as in terms of $SU(2)_L \otimes SU(2)_R \otimes SU(4)_c$ $((2,1,4) \oplus (1,2,4)$: right table).

ν_L	$d_R{}^c$	$d_G{}^c$	$d_B{}^c$	u_R	u_G	u_B	e^+
e^-	$u_R{}^c$	$u_G{}^c$	$u_B{}^c$	d_R	d_G	d_B	$\nu^c{}_L$
5^*			10				1

ν_L	u_R	u_G	u_B	$SU(2)_L$
e^-	d_R	d_G	d_B	
e^+	$d_R{}^c$	$d_G{}^c$	$d_B{}^c$	$SU(2)_R$
$\nu^c{}_L$	$u_R{}^c$	$u_G{}^c$	$u_B{}^c$	
$SU(4)_c$				

The 16 fermions contain the known 15 fermions and a right-handed singlet neutrino.

The third feature is that this $\nu_L{}^c = \nu_R$ is a candidate for the partner in the seesaw mechanism, offering a possibility to solve the neutrino mass problem.

The fourth attractive feature is that it provides more than one route to break the symmetry. It provides an extra handle which can be used to adjust the evolution of the running coupling constant minutely. For instance, it is possible to have the three constants meet together and to reproduce the observed Weinberg angle as well as the proton decay rate.

Thus the $SO(10)$ model has many attractive features not provided by other models and is one among the favorites of many theorists in discussing the GUT model.

3.4.2
New Gauge Bosons Z' and W'

When the GUT group G contains a subgroup other than $SU(3) \times SU(2) \times U(1)$, it means that extra gauge bosons and with them associated gauge interactions exist. For instance, the left–right symmetric model contains $SU(2)_R$ and induces new gauge bosons $W_R^\pm$, W_R^0. Phenomenologically, the FCNC (flavor changing neutral current which affects $K^0 - \overline{K}^0$ mixing, etc.) imposes a constraint on the mass of W_R and its mass must be larger than a few times that of the SM W boson (see discussions in Section 2.1.2). Most GUT symmetry groups contain extra $U(1)$ in addition to the SM $U(1)_Y$, and hence extra neutral gauge bosons. We denote them as $U(1)_E$ (E = extra) and Z', respectively. Then the Lagrangian to describe the neutral current interaction can be expressed as

$$-\mathscr{L}_{NC} = eJ^\mu_{\text{EM}} A_\mu + g_z J^\mu_{Z1} Z^0_{1\mu} + g_E J^\mu_{Z2} Z^0_{2\mu} \tag{3.66}$$

where A_μ, Z_1^0, J^μ_{Z1} are the photon, Z^0, and the neutral current in the SM. They are expressed as

$$g_z = \frac{g_W}{\cos\hat{\theta}_W} = \frac{e}{\sin\hat{\theta}_W \cos\hat{\theta}_W} \tag{3.67a}$$

$$J^\mu_{Z1} = \sum_f \overline{f}\gamma^\mu \left[\varepsilon^0_{L1}(f)\frac{1-\gamma^5}{2} + \varepsilon^0_{R1}(f)\frac{1+\gamma^5}{2}\right] f \tag{3.67b}$$

$$\varepsilon^0_{L1}(f) = I_{3f} - Q_f \sin^2\hat{\theta}_W, \quad \varepsilon^0_{L1}(f) = -Q_f \sin^2\hat{\theta}_W, \tag{3.67c}$$

Here, f denotes a fermion field and I_{3f}, Q_f are its isospin third component and electric charge. The third term in Eq. (3.66) is a new interaction associated with the extra boson Z_2^0. Its associated neutral current is written as

$$J_{Z2}^{\mu} = \sum_f \bar{f}\gamma^{\mu}\left[\varepsilon_{L2}^0(f)\frac{1-\gamma^5}{2} + \varepsilon_{R2}^0(f)\frac{1+\gamma^5}{2}\right]f$$

$$= \sum_f \left[Y(f_L)\bar{f}_L\gamma^{\mu}f_L + Y(f_L{}^c)\overline{f_L{}^c}\gamma^{\mu}f_L{}^c\right] \tag{3.68a}$$

$$Y(f_L) = \varepsilon_{L2}^0(f), \qquad Y(f_L{}^c) = -\varepsilon_{R2}^0(f) \tag{3.68b}$$

In deriving the last equality, we used the relation $\overline{\psi}_R\gamma^{\mu}\psi_R = -\overline{(\psi^c)_L}\gamma^{\mu}\psi_L^c$. $Y(f_L)$, $Y(f_L{}^c)$ are hypercharges associated with the group $U(1)_E$ that the chiral fermions f_L,f_L^c posses.

As Z_1^0, Z_2^0 couple to the same fermion, they mix and, as a result, the physical mass eigenstates are expressed as

$$Z_1 = Z_1^0\cos\theta_E + Z_2^0\sin\theta_E \tag{3.69a}$$

$$Z_2 = -Z_1^0\sin\theta_E + Z_2^0\cos\theta_E \tag{3.69b}$$

We interpret that the observed Z^0 is Z_1 and Z_2 is the newly introduced Z'. The coupling strengths of the new Z_1, Z_2 are given by

$$g_z\varepsilon_{L1}(f) = g_z\varepsilon_{L1}^0(f)\cos\theta_E + g_E\varepsilon_{L2}^0(f)\sin\theta_E \tag{3.70a}$$

$$g_z\varepsilon_{L2}(f) = -g_z\varepsilon_{L1}^0(f)\sin\theta_E + g_E\varepsilon_{L2}^0(f)\cos\theta_E \tag{3.70b}$$

The coupling strength to the right-handed fermions f_R can be obtained by replacing $L \to R$ in the above expression. The corresponding mass matrix is written as

$$\tilde{M} = \begin{bmatrix} M_0^2 & \Delta^2 \\ \Delta^2 & M_E^2 \end{bmatrix}, \qquad M_0 = \frac{m_W}{\cos\hat{\theta}_W} \tag{3.71}$$

M_0, M_E are the masses of Z_1^0, Z_2^0 before mixing. Masses of Z_1, Z_2, the mixing angle, and the nondiagonal elements can be expressed as

$$M_{1,2}^2 = \frac{1}{2}\left[M_0^2 + M_E^2 \pm \sqrt{(M_0^2 - M_E^2)^2 + 4\Delta^2}\right] \tag{3.72a}$$

$$\tan 2\theta_E = \frac{2\Delta^2}{M_0^2 - M_E^2} \tag{3.72b}$$

Alternatively, the mixing angle, and so on, can be expressed in terms of physical masses.

$$\tan^2\theta_E = \frac{M_0^2 - M_1^2}{M_2^2 - M_0^2} \tag{3.73a}$$

$$\Delta^2 = \sqrt{(M_2^2 - M_0^2)(M_0^2 - M_1^2)} \tag{3.73b}$$

From an inspection of the above-mentioned formulas, we find that the SM is modified as follows:

1) The coupling constant of the SM Z boson changes.
2) The mass of the SM Z boson becomes smaller.
3) New processes to exchange Z' appear.

As reproducibility of the SM is excellent, we can approximate $M_1 = M_0 - \delta$, $\delta \ll M_1 \approx M_0 \ll M_2$. Then the mixing angle becomes

$$\theta_E \approx \frac{\Delta^2}{M_2^2} \approx O\left(\frac{\sqrt{2M_0\delta}}{M_2}\right) \tag{3.74}$$

Thus corrections to the known mass and the coupling constant are small.

The coupling constant g_E depends on what group one chooses. Among the many models proposed so far, we list only a few representative ones in the following. A favorite model is E_6, which is inspired by the superstring model. Depending on how one breaks the symmetry, various kind of neural gauge bosons Z_ψ, Z_χ, Z_η appear.

$$\begin{aligned} E_6 \to\ & SO(10) \times U(1)_\psi \\ & \to SU(5) \times U(1)_\chi \times U(1)_\psi \\ & \to SU(3) \times SU(2)_L \times U(1)_Y \times U(1)_\chi \times U(1)_\psi \\ & \to SU(3) \times SU(2)_L \times U(1)_Y \times U(1)_\eta \end{aligned} \tag{3.75}$$

Using an alternative expression

$$E_6 \to G' \times U(1)_\beta, \qquad G' \supset SU(3) \times SU(2) \times U(1)_Y, \tag{3.76}$$

we can write generically

$$Z_2^0 = Z_\beta \equiv \cos\beta Z_\chi + \sin\beta Z_\psi \tag{3.77}$$

The coupling strength of Z_β to fermions in the fundamental representation **27** of E_6 is proportional to the hypercharge Y_β defined by Eq. (3.68b). Z_χ, Z_ψ correspond to $\beta = 0, \pi/2$ and $Z_\eta = -Z(\beta = \pi - \tan^{-1}\sqrt{5/3})$. Then the ratio of the coupling constant can be expressed as

$$\frac{g_E}{g_z} = \sqrt{\frac{5}{3}}\lambda \sin\hat{\theta}_W \tag{3.78}$$

with λ representing a numerical factor dependent on the model and the energy scale. At the GUT energy scale, $\tan\hat{\theta}_W = \sqrt{3/5}$. In Table 3.6, we list the hypercharge of the fermions belonging to the E_6 fundamental representation which can be decomposed to $\mathbf{16} \oplus \mathbf{10} \oplus \mathbf{1}$ of $SO(10)$ [136, 378].

Global analyses including the contribution of the extra Z' boson tend to reduce R_{ll} and increase R[5] for the processes $e^-e^+ \to l\bar{l}, e^-e^+ \to q\bar{q}$ compared to the SM

---

[5]

$$R_{ll} = \frac{\sigma(e\bar{e} \to l\bar{l})}{\sigma(e\bar{e} \to \mu\bar{\mu})}, \qquad R = \frac{\sigma(e\bar{e} \to \text{hadrons})}{\sigma(e\bar{e} \to \mu\bar{\mu})} \tag{3.79}$$

Table 3.6 Hypercharge of fermions in various models evolved from E_6 [136].

$SO(10)$	$SU(5)$	$\sqrt{24}Y_\psi$	$2\sqrt{10}Y_\chi$	$2\sqrt{15}Y_\eta$
16	$10\,(u, d, u^c, e^+)_L$	1	−1	−2
	$5^*(d^c, \nu, e^-)_L$	1	3	1
	$1\quad \overline{N}_L$	1	−5	−5
10	$5\,(D, \overline{E}^0, E^+)_L$	−2	2	4
	$5^*(\overline{D}, E^0, E^-)_L$	−2	−2	1
1	S_L^0	4	0	−5

$\overline{N}, D, E, S$ are newly added exotic fermions (or super partner of the Higgs in the SUSY GUT).

Table 3.7 Ninety-five percent CL lower limits on various extra Z' boson masses (GeV) and 90% CL ranges for mixing angle $\sin\theta_E$ from electroweak precision data (columns 2 and 3) [7, 374, 379, 380].

Z'	$M_{Z'}$/EWPD	$\sin\theta_E \times 10^{-3}$	$M_{Z'}$ Tevatron	$M_{Z'}$ LEP2
χ	551	−2.0 to +1.5	822	673
ψ	151	−1.3 to +2.4	822	481
η	379	−6.2 to +1.1	891	434
LR	570	−0.9 to +1.7	630	804
SSM	822	−4.1 to +0.3	923	1787

Tevatron and LEP2 searches for direct production (columns 4 and 5). LR means left–right symmetric model where symmetry-breaking path is $SO(10) \times U(1)_\psi \to SU(4) \times SU(2)_L \times SU(2)_R \times U(1)_\psi$. Here the combination of third component of the isospin T_{3R} and Y_ψ generate the SM hypercharge and extra Z boson LR. SSM (sequential standard model) means an extra Z' is assumed to have the same coupling as the SM ($\lambda = 1$). The last column contains LEPII results, which analyzed the precision production cross section of $q\overline{q}, l\overline{l}$ including the effective four-Fermi interaction due to Z'.

analysis. Note that low-energy neutrino scattering and $e\overline{e}$ reactions are sensitive to the mass of Z', and $e\overline{e}$ processes at Z resonance are sensitive to the $Z - Z'$ mixing. A result of global analysis using pre-LHC data is given in Table 3.7 [136].

Direct Searches for the Z' As the Z' has exactly the same quantum number as γ or Z of the SM, it can be produced in the e^-e^+ or $p\overline{p}$ collider. The cross section for colliding hadrons A and B at center-of-mass energy $\sqrt{s}$ to produce a Z_α of mass M_α can be obtained from that of the Drell–Yan process by setting the invariant mass of the muon pair at M_α [see Eq. (3.20)–(3.24) in [2]]. Neglecting the fermion mass, and in the narrow width approximation, the cross section to produce Z_α with rapidity y can be expressed as

$$\frac{d\sigma}{dy} = \frac{4\pi^2 x_a x_b}{3M_\alpha^3} \sum_a \left[f_A^{q_a}(x_1) f_B^{\overline{q}_a}(x_2) + f_A^{\overline{q}_a}(x_1) f_B^{q_a}(x_2)\right] \Gamma(Z_\alpha \to q_a \overline{q}_a) \tag{3.80}$$

where $f_{A,B}^{q_a,\bar{q}_a}$ are the quark (or antiquark) distribution functions in hadrons A or B. The momentum fractions of the quarks relative to their parent protons are given by

$$x_a = \frac{M_\alpha}{\sqrt{s}} e^y, \quad x_b = \frac{M_\alpha}{\sqrt{s}} e^{-y} \tag{3.81}$$

The decay width can be obtained from that of Z_{SM} [see Eqs.(4.3) and (4.6) of [2]] by replacing appropriate parameters.

$$\Gamma(Z_{\text{SM}} \to f\bar{f}) = \frac{N_c^f g_z^2 m_Z}{48\pi}(v_f^2 + a_f^2) = \frac{N_c^f g_z^2 m_Z}{24\pi}\left[\varepsilon_L(f)^2 + \varepsilon_R(f)^2\right] \tag{3.82a}$$

$$\to \quad \Gamma(Z_\alpha \to f\bar{f}) = \frac{N_c^f g_\alpha^2 M_\alpha}{24\pi}\left[\varepsilon_L^\alpha(f)^2 + \varepsilon_R^\alpha(f)^2\right] \tag{3.82b}$$

where $N_c^f(= 3, 1$ for q, $l)$ is the color factor. For the actual calculation, one further takes into account of mixing effects. Phenomenologically, Eq. (3.80) is often replaced in good approximation by [136, 379]

$$\sigma_{Z'} = \frac{1}{s} c_{Z'} C\exp\left(-A\frac{M_{Z'}}{\sqrt{s}}\right) \tag{3.83}$$

where $C = 600(300)$ and $A = 32(20)$ for pp($p\bar{p}$) collisions and $c_{Z'}$ is a model-dependent factor.

The detection signal for Z' production is a resonance peak in the spectrum of the lepton pair invariant mass in the process $pp(p\bar{p}) \to l\bar{l}$, $l = e, \mu$. Direct productions of Z' ($p + p \to Z' + X \to l\bar{l} + X'$) at LHC give much better lower limits to the Z' masses than the global analysis of the precision data. Figure 3.6 depicts an example of dielectron spectrum (a) and mass limits compared with various theoretical models (b) [381].

The present lower limit of the Z' mass is ~2.5 TeV.

A New Gauge Boson W'

W_R Searches at Hadron Colliders W_R appears not only in the left–right symmetric model but also in a variety of new extended models beyond the SM if it includes additional $SU(2)$ symmetry. This happens, for example, for the little Higgs model we mentioned in Section 1.5.2. Direct production of W_R, Z_R with much higher mass can be achieved at the LHC. W_R can decay to a right-handed neutrino via $W_R \to l_R \nu_R$, $l = e, \mu\,\tau$ if $m(\nu_R) < m(W_R)$. However, the masses of the right-handed neutrino are not constrained as far as they do not couple to the SM particles. $m(\nu_R) > m(W_R)$, the limit obtained from leptonic decays does not apply. In this case, decays $W_R \to qq' \to 2$ jets have to be analyzed. However, if $m(W_R) > 180$ GeV, the process $W_R \to t_R b_R \to (b_L W_L) b_R$ is possible because of $t_R - t_L$ mixing due to the top's large mass. The above consideration means that $W_R \to l + X$ is a good channel to look for W_R first. If a signal is detected, one has to perform additional

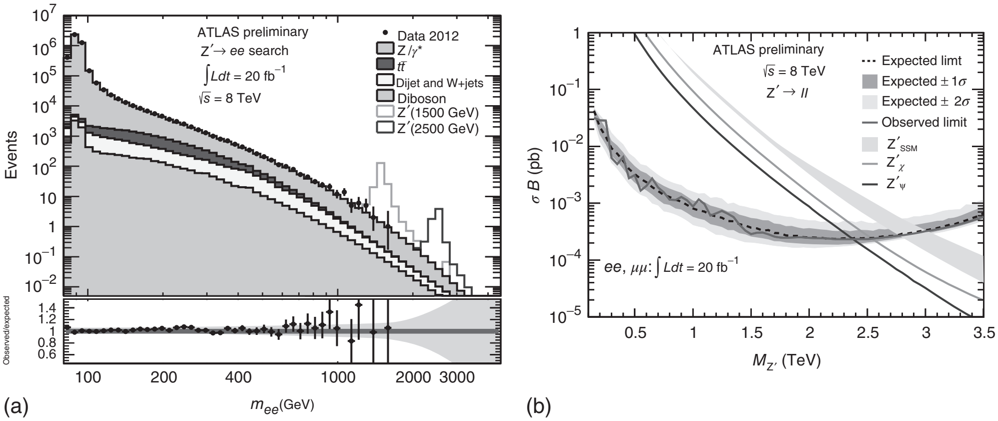


Figure 3.6 (a) Dielectron invariant mass (m_{ee}) distributions, compared with the stacked sum of all expected backgrounds, with two example Z'_{SSM} signals overlaid. Z'_{SSM} is the sequential Z' which is a heavy Z with the SM coupling. The bin width is constant in $\log m_{ee}$. (Bottom inset) The black points show the ratio of observed to expected events with statistical uncertainty, while the shaded band indicates the mass-dependent systematic uncertainty on the sum of the backgrounds. (b) Expected and observed 95% CL limits on $\sigma \cdot B$ and the expected $\sigma \cdot B$ for Z'_{SSM} production and the two E_6-motivated Z' models (Z'_χ, Z'_ψ) with lowest and highest $\sigma \cdot B$ for the combination of the dielectron and dimuon channels. The dashed lines around the Z'_{SSM} theory curve represent the theoretical uncertainty, which is similar for the other theory curves. (Reproduced with permission of [381–385].)

experiment to distinguish W_R from other W', for example, model-dependent analyses of $W' \to WZ$ decay, but with null results the limit can be applied to either of them.

New heavily charged gauge bosons including W_R are generically referred to as W'. If W' is left-handed with the same coupling as the SM W, it is referred to as the W(SSM) (*sequential standard model*). In the extra-dimensional models, a sequence of left-handed W bosons will appear if the $SU(2)$ particles live in the bulk. The left-handed W' can interfere with the SM W constructively or destructively depending on the sign of the coupling. The effect may contribute as much as 5–20%. If we neglect the interference effect, W' and W_R produce the same signature in the fermionic decay channels. W'/W_R can be searched in the transverse mass distribution of the $l - \nu$ system, which is defined by

$$M_T \equiv \sqrt{2p_T^l E_T^{\text{miss}}(1 - \cos\Delta\phi_{l\nu})} \tag{3.84}$$

where $\Delta\phi_{l\nu}$ is the azimuthal opening angle between the charged transverse momentum (p_T^l) and the missing transverse energy (E_T^{miss}) direction. The neutrino is not detected directly, but gives rise to experimentally observable E_T^{miss}. The M_T

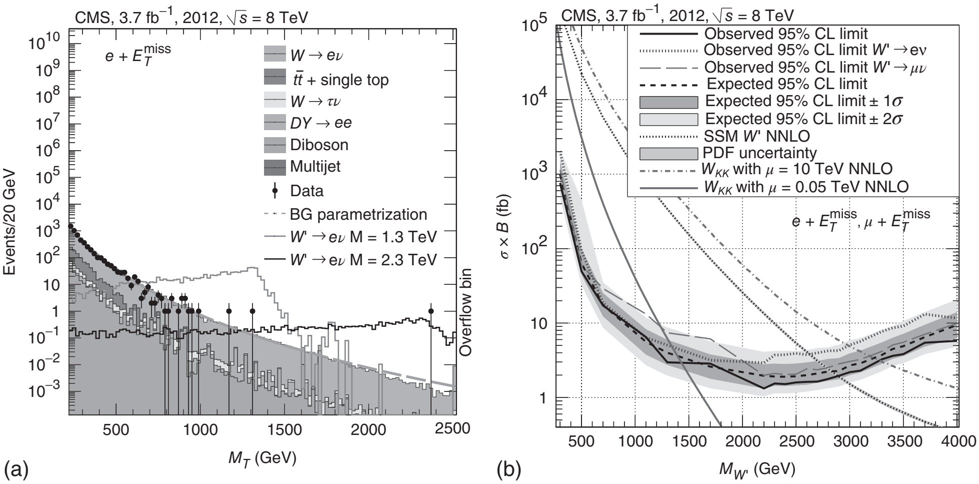


Figure 3.7 (a) Observed transverse mass distributions for the electron channels. Simulated signal distributions for a W' are also shown, including detector resolution effects. The simulated background labeled as "diboson" includes WW, ZZ, and WZ contributions. The top background prediction includes single top and top pair production. The total background prediction from a fit to the simulated transverse mass spectrum in each channel is shown by the dashed line. (b) Limits on the cross section times the single channel-branching fraction ($\sigma \cdot B$) for heavy W' bosons for the electron and the muon channels. For the individual channels, only the observed limits are shown. For the combination, the observed limit, the expected limit, the 1σ, and the 2σ bands are displayed. All limits are displayed for the single channel-branching fraction. The W' mass limits are for the models of a SSM W' and $W_{KK}^{n=2}$ in split UED (universal extra dimension model) (see Figure 6.2 for interpretation of extra dimensional models). (Reproduced with permission of [386–388].) (Please find a color version of this figure on the color plates.)

distribution of a two-body decay exhibits a sharp kinematical cut, referred to as the *Jacobian peak*. Figure 3.7a shows the $M_T(l\nu)$ distribution for $l = e$ collected by the CMS group [386]. A hypothetical spectrum, if W_R with $m(W_R) = 1.3$ TeV or 2.3 TeV exists, is also shown in the figure. The observed data can be reproduced well with the SM, and no sign of W'/W_R was detected. The obtained upper limit was (Figure 3.7b)

$$m(W') > 2.5\ \text{TeV} \quad 95\%\text{CL} \tag{3.85}$$

Later, it was further improved to $m(W') > 2.9$ TeV.

Searches in the $W_R\,(W'_L) \to tb$ channel give a smaller lower limit of 1.85 (1.51) TeV at the 95% CL [388]. Figure 3.8 shows the results of new resonance searches in various models carried out by the CMS group. Similar results were also obtained by the ATLAS group [389].

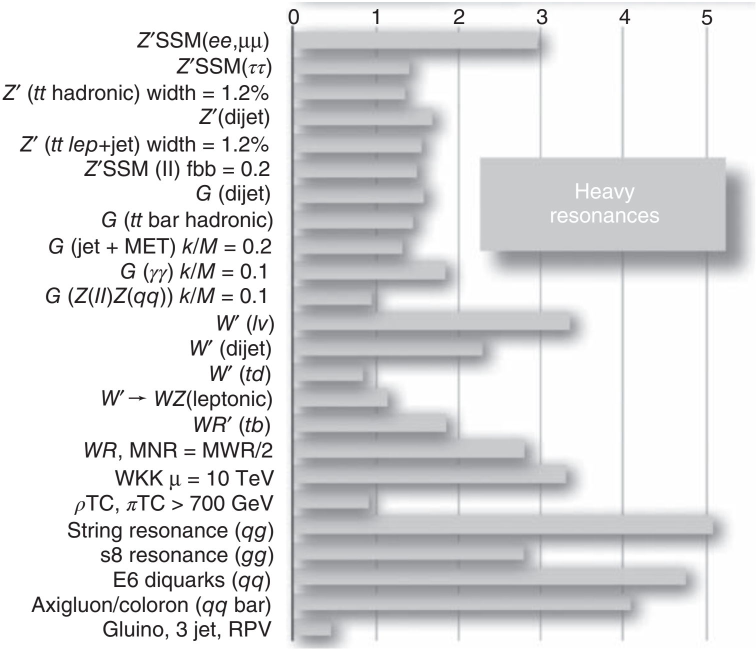


Figure 3.8 Upper limits of heavy resonances obtained by CMS circa 2013. (Reproduced with permission of [390].)

3.5
Hierarchy Problem

In the GUTs, two inherent energy scales coexist: the GUT scale ($<\Phi> \sim 10^{15}$ GeV), and the electroweak scale ($<H> \sim (\sqrt{2}G_F)^{-1/2} = 250$ GeV. If the notion of the grand unification is correct, we need to construct a theoretical framework to treat phenomena of two different energy scales differing by an order $\sim 10^{-13}$. This is the (so-called gauge or big) hierarchy problem. We do not have a satisfactory answer to the fundamental question why such a large hierarchy exists.6) But as a result of it, we encounter a technical problem of "fine-tuning." Let us consider the Higgs mass again. The following arguments overlap somewhat with those given in Section 1.1.3 but here we are discussing the big hierarchy problem. In the tree approximation, it is given by

$$m_{H^0}^2 = -2\mu^2 = 4\lambda <H_5>^2 \tag{3.86}$$

Contributions of the radiative processes given in Figure 3.9a–c are given by

$$\delta m_{H^0}^2 \sim Cg^2\Lambda^2 \tag{3.87}$$

where C is a constant, g the coupling strength, and Λ a cut-off energy scale.

---

6) Some extra-dimensional models provide a possible answer. See Chapter 6.

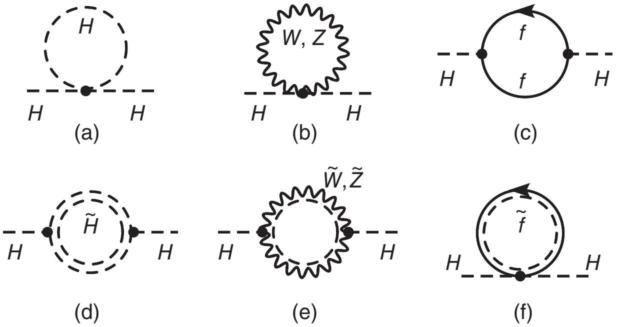


Figure 3.9 Radiative corrections to the Higgs mass. (a)–(c) are *StandardModel* corrections, (d)–(f) are the corresponding SUSY corrections.

The cutoff exhibits the degree of divergence in the renormalization. An extraordinary feature of the Higgs (or scalar boson in general) mass radiative correction is that it is proportional to the square of the cutoff. This is in sharp contrast to those of the gauge boson or fermion, which are only logarithmic. This is because they are protected by symmetries, the gauge boson by the gauge symmetry and the fermion by chiral symmetry. There is no such symmetry to protect the divergence of the scalar meson, and the quadratic divergence is one of the characteristic features.

In the GUTs, an interaction between H and Φ exists (like $\mathscr{L}_{\text{INT}} \sim \Phi^2 H^2$), which arises as a result of processes depicted in Figure 3.10.

Therefore, processes including Φ loops have also to be included in the radiative corrections. The correction also diverges quadratically as a function of the cutoff, but this time the cutoff is of the GUT scale and the mass correction is $\delta m^2_{H^0} \sim O(\Lambda^2) \sim m(\Phi)^2 \sim 10^{30}\ \text{GeV}^2$. As the Higgs mass is $m^2_{H^0} \simeq v^2 \simeq (250\,\text{GeV})^2$, one has to compensate the GUT correction with accuracy to the twenty-sixth order of magnitude. The fine-tuning has to be carried out at every order of perturbation series. This is logically very unnatural. The fine-tuning problem is inherent to the GUTs and is model-independent. The problem has to be solved fundamentally. Generally speaking, there are three ways to do it.

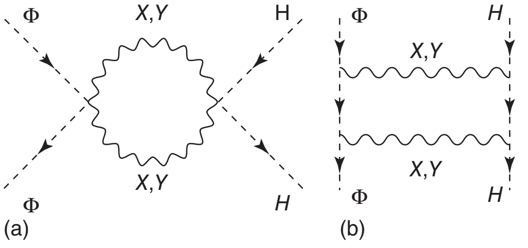


Figure 3.10 Examples of graphs inducing $\Phi\Phi HH$ interactions. This kind of interaction, even if nonexistent in the original Lagrangian, always appears as radiative corrections. The GUT gauge boson X, Y can couple to both H and Φ, and processes exchanging two of them (XX, XY, YY) generate the interaction effectively.

1) **Composite scalar** The first solution is to assume that no elementary Higgs exists. It is analogous to superconductivity, which served as a model for the Higgs mechanism. There, the Cooper pair plays the role of the Higgs field and induces the spontaneous symmetry breakdown. Therefore, we may assume that the Higgs is a composite of a fermion pair. If it is true, the constituent particles tend to fly apart to become free when the energy deposited on them is much higher than the binding energy. The Cooper pair dissociates at high temperature. Technically, it is described as suppression of the form factor to damp the divergence [see Section 17.2 of [1]]. If one ascribes the role to the known fermions, it is the top condensate model [46, 48, 391–393] but new possibilities arise if one ascribes the role to new fermions. A representative model is the technicolor model described in Section 1.5.1. One assumes the existence of more fundamental fermions referred to as *techni-fermions* which interact under the technicolor force. The technicolor is essentially a scaled-up version of QCD and the particle spectra are very similar with masses ~ × 1000 larger. In the technicolor model, the techni-pion is assumed to be the Higgs in the SM. The technicolor model has a rich content and is a very attractive model conceptually. However, viable models that can reproduce phenomenology are not yet available and early enthusiasm is somewhat subdued. Yet the basic notion of the concept could well be effective. A recent variation of the composite model includes the little Higgs model though it does not address to the big hierarchy problem.
2) **Extra dimension** Enlarging the space dimension from three to $D = 3 + \delta$ ($\delta = 1, 2, \ldots$) provides another possibility to solve the fine-tuning problem. It might even answer the fundamental question of why such hierarchy exists by claiming that no big hierarchy exists. We defer the discussion to Chapter 6.
3) **Supersymmetry** The third solution is to introduce new particles whose role is to compensate the divergence of loop integrals piece by piece. Radiative corrections by fermion loops have opposite sign to those by bosons owing to the Fermi–Dirac statistics. If there exist new fermions corresponding to known bosons with the same mass and the same coupling, they can compensate the boson loop divergence. Consider $\tilde{H}, \tilde{W}, \tilde{Z}$, which are fermions with the same mass and the same coupling as H, W, Z. Their loop contributions (Figure3.9d,e) can compensate the divergence of the Higgs and W/Z boson loops in Figure 3.9a,b. The SM fermion loop contribution Figure 3.9c can be compensated by considering a corresponding scalar boson $\tilde{f}$ (Figure 3.9f). They amount to a requirement of a symmetry under exchange of bosons and fermions, which is referred to as *supersymmetry*. If the coupling strength is the same, the mass correction is of the order

$$\delta m_H^2 \sim Cg^2\Big(\Lambda^2 + O(m_B^2)\Big) - Cg^2\Big(\Lambda^2 + O(m_F^2)\Big) \sim Cg^2(m_B^2 - m_F^2) \quad (3.88)$$

where m_B, m_F are the masses of the particles to be paired. If the symmetry is exact, the cancellation is complete and there are no radiative corrections.

But even if the symmetry is broken, it still can solve the fine-tuning problem provided the mass difference of the partners remain on the order $\lesssim 1$ TeV [394].

If one worries that the method is *ad hoc* and too artificial, it is a historically proved valid prescription. Consider, for example, the role of the charm quark in the GIM (Glashow-Iliopoulos-Maiani) mechanism. It was introduced to suppress the flavor-changing neutral current. The compensation would have been exact if the mass and the coupling were the same as those of the u quark. The adjusted mass to reproduce the small residual effect (mass difference of $K_S^0 - K_L^0$) reproduced the observed charm mass approximately. Also, notice that the road from the four-Fermi interaction to the SM followed this route. By introducing new particles to cancel divergences induced by massive gauge bosons, one is led to the spontaneously broken symmetry [395–398] (see arguments in Section 1.7 of [2]).

Anyway, accommodating the SUSY in the framework of the GUTs (referred to as *SUSY GUTs*) turned out to solve the problem of the Weinberg angle and the proton decay lifetime, which we discussed before. We will conclude this chapter by discussing how these problems can be solved by using the SUSY GUTs. As details of the SUSY is treated in the next chapter, the reader may refer to it for terminologies and rules cited here without proof.

3.6
SUSY GUT

Phenomenologically, the SUSY GUT is almost the same as the MSSM (minimum supersymmetric extension of the SM) except for some GUT-specific predictions. Let us see how it can improve the non-SUSY GUTs.

Gauge Unification and the Weinberg Angle The SUSY, if it exists, is broken. But if it is restored at the energy scale beyond ~1 TeV, it does not contradict the observations and nicely solves the quadratic divergence problem encountered in the Higgs mass corrections. If the supersymmetric partner of the SM particles (referred to as *sparticles*) exist, they have to be included in the RGEs for evolution of the coupling constants above the energy scale larger than their mass values. Therefore, the beta functions we discussed in Eq. (3.51) have to be modified as follows [399]:

$$\begin{aligned}
\frac{1}{\alpha_i(\mu)} &= \frac{1}{\alpha_{\text{GUT}}} + \frac{\beta_i}{2\pi}\ln\left(\frac{\mu}{M_{\text{GUT}}}\right) + \cdots \\
\beta_3 &= \frac{1}{3}\left[(33-4n_g)-(6+2n_g)\right] \\
\beta_2 &= \frac{1}{3}\left[\left(22-4n_g-\frac{1}{2}\right)-\left(4+2n_g+\frac{5}{2}\right)\right] \\
\beta_1 &= \frac{1}{3}\left[\left(-4n_g-\frac{3}{10}\right)-\left(2n_g+\frac{3}{2}\right)\right]
\end{aligned} \tag{3.89}$$

The first terms on the right-hand side of the β_i function are the beta functions that already appeared in the $SU(5)$ GUT, and the second terms represent additional SUSY contributions. n_g is the number of generations set to 3. Values of the beta functions are

$$(\beta_1, \beta_2, \beta_3) = \begin{cases} (-41/10,\ 19/6,\ 7) & \text{non-SUSY GUT SU(5)} \\ (-33/5,\ -1,\ 3) & \text{SUSY GUT SU(5)} \end{cases} \tag{3.90}$$

The SUSY term has the effect of ameliorating the evolving speed of the RGEs differently for each gauge group. An analysis of the LEP precision data using the RGEs including the NLO corrections [399] shows that there are no unification points in the conventional (non-SUSY) $SU(5)$ GUT, but in SUSY $SU(5)$ all the coupling constants meet at $M_{ss} \sim 2 \times 10^{16}$ GeV (see Figure 3.11).

Accordingly, the evolution of the Weinberg angle (see Eq. (3.52)) is also modified to give [401]

$$\sin^2 \hat{\theta}_W = \frac{3}{8} - \frac{15}{8\pi}\alpha_{\mathrm{EM}} \ln\left(\frac{M_{ss}}{m_W}\right) \simeq 0.23 \tag{3.91}$$

which reproduces the precise experimental data. Furthermore, α_{GUT} which is recalculated using SUSY modified Eq. (3.51) gives

$$\alpha_{\mathrm{GUT}} \simeq 1/25 = 0.04 \tag{3.92}$$

and differs from the value of $\alpha_5 \simeq 0.025$ [Eq. (3.53)] obtained form non-SUSY $SU(5)$.

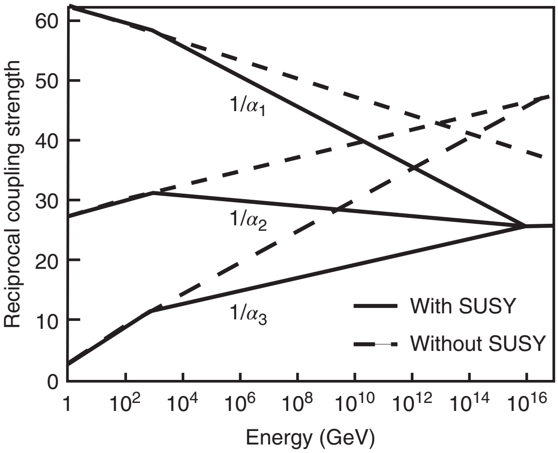


Figure 3.11 Unification of electroweak and strong forces. Plotted are the inverse of $\alpha_1, \alpha_2, \alpha_3$, which are the coupling strengths of $U(1)$, $SU(2)$, $SU(3)$ gauge symmetries. The couplings do not meet at a point without supersymmetry, but they do with supersymmetry. The supersymmetry is assumed to be broken below ~1 TeV, which is the cause of kinks in the figure. (Reproduced with permission of [399, 400].)

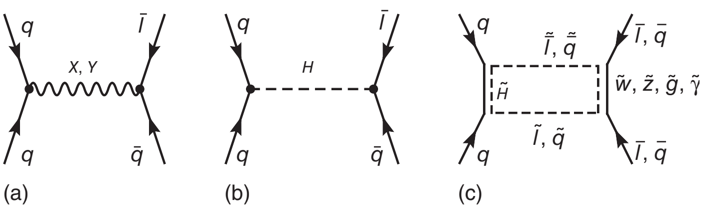


Figure 3.12 Proton decay by SUSY-$SU(5)$. (a) Ordinary $SU(5)$. (b) Higgs intermediate state. (c) Higgsino/squark/gaugino intermediate states.

Proton Decay The unification energy M_{ss} being larger than $M_{\rm GUT}$ determined from non-SUSY $SU(5)$, the masses of X, Y also become larger. As a result, the lifetime of the proton $\tau_{p,ss}$ (we denote the variables of SUSY GUTs as ss and those of non-SUSY GUT as *ns*) is also modified to

$$\tau_{p,ss} \approx \left(\frac{M_{X,ss}}{M_{X,ns}}\right)^4 \tau_{p,ns} \approx 10^{38\pm 2} \text{ years} \tag{3.93}$$

which is a value that does not contradict the observations. This value is unrealizable with current detectors. In the SUSY GUT, however, one has to consider the contribution of extra particles, namely that of Higgs (H_{SU5}) and higgsino ($\tilde{H}^0$), as shown in Figure 3.12.

The contribution of the Higgs exchange process is proportional to the mass it couples, and hence the process Figure 3.12b is negligible unless one makes M_H as large as $\sim 10^{11}$ GeV. The higgsino exchange, on the other hand, has a compensating enhancement factor because it is a fermion. Its amplitude is proportional to $M_{\tilde{H}^0}^{-1}$, which is in contrast to $\sim M_X^{-2}$ of the X boson. Ambiguities due to the interplay between these compensating parameters make an accurate estimation difficult, but with reasonable parameter settings it is the dominant mode and the predicted value is within the experimental reach [402, 403]. As the higgsino couples to the heaviest particle, the dominant decay mode is

$$p \to K^+\nu \tag{3.94}$$

which is in sharp contrast to $e\pi^0$, $\nu\pi^+$ dominance in the non-SUSY GUTs and is a good signature to discriminate the underlying mechanism if it is due to SUSY or non-SUSY GUT. Recent Super-Kamiokande bounds on partial lives are [404]

$$\tau(p \to K^+\nu) > 2.3 \times 10^{33} \text{ years}, \qquad \tau(p \to K^0 e^+) > 2.0 \times 10^{33} \text{ years} \tag{3.95}$$

The proton decay data is reaching the limit that non-SUSY GUTs model can allow. If the SUSY GUTs are correct, it should be observed soon.

In summary, the SUSY GUTs are consistent with the experimental data and theoretically an attractive option as candidates for the grand unification.

4
Supersymmetry I: Basics

4.1
Introduction

The supersymmetry (SUSY) unifies fermions and bosons.* The corresponding operator converts fermion fields to boson field and vice versa. Such transformations are counterintuitive because of the very different characteristics attached to both fields. The fermion respects the Pauli exclusion principle and its number is a conserved quantity. All the special features that are ascribed to the fermion seem incompatible with those of the boson. Therefore, it was revolutionary when it was proposed. Nowadays, it is considered as an indispensable symmetry. It provides a solution to the big as well as the little hierarchy, and a way to quantize the gravity. The superstring model makes extensive use of the SUSY in constructing the unified theory of everything.

The SUSY is different from all other symmetries in that its operator is fermionic. Familiar operators such as the Lagrangian, the energy–momentum P^{μ}, and the angular momentum $M^{\mu\nu}$ are a Lorentz scalar, vector, or tensor, and internal symmetry operators such as the isospin operator I are scalars. These operators are generally polynomials of field operators. Let us remind you that the field operator includes spinors and tensors. Then, one might also think that there could exist a symmetry operator in the spinor form. The SUSY operator $Q_{\alpha i}$ is exactly that. Here, α denotes the spinor index and $i\,(1 \leq i \leq N)$ denotes the internal degrees of freedom. If one makes the spinor operator act on a bosonic state $|B>$ which is a vacuum state multiplied by a bosonic operator, the product transforms as a spinor; hence, it represents a fermionic state $|F>$. That is,

$$Q_{\alpha i}|B>\sim |F>, \qquad Q_{\alpha i}|F>\sim |B> \tag{4.1}$$

Therefore, the primary effect of the SUSY operator is to make a bridge between the boson and the fermion.

---

* This chapter is provided to learn basic algebra and to derive a Lagrangian of SUSY. Those who are only interested in the phenomenology of the SUSY can skip this chapter except the introduction section and go directly to the next chapter.

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

4.1.1
Toy Model

As an illustration, let us consider a toy model [363], a harmonic oscillator consisting of a fermion and a boson. Writing their annihilation operators as *'a'* (fermion) and *'b'* (boson), the simplest Hamiltonian in quantum mechanics is expressed as

$$H = \omega_F a^\dagger a + \omega_B b^\dagger b \tag{4.2}$$

$$\{a,\, a^\dagger\} = [b,\, b^\dagger] = 1, \quad \{a,\, a\} = \{a^\dagger,\, a^\dagger\} = [b,\, b] = [b^\dagger,\, b] = 0 \tag{4.3}$$

where the parentheses mean $\{A, B\} = AB + BA$, $[A, B] = AB - BA$. Defining a SUSY operator Q, $Q^\dagger$ by

$$Q = b^\dagger a, \quad Q^\dagger = a^\dagger b, \tag{4.4}$$

they satisfy relations

$$[Q,\, H] = (\omega_F - \omega_B)Q, \quad [Q^\dagger,\, H] = -(\omega_F - \omega_B)Q^\dagger \tag{4.5}$$

Therefore, if $\omega_F = \omega_B$, Q, $Q^\dagger$ commute with the Hamiltonian and are conserved operators. They also satisfy

$$[N_B,\, Q] = [b^\dagger b,\, b^\dagger a] = \ b^\dagger a = \ Q, \quad [N_B,\, Q^\dagger] = -Q^\dagger \tag{4.6a}$$

$$[N_F,\, Q] = [a^\dagger a, b^\dagger a] = -b^\dagger a = -Q, \quad [N_F,\, Q^\dagger] = Q^\dagger \tag{4.6b}$$

Therefore, when the operators Q, $Q^\dagger$ act on states, they change the bosonic or fermionic numbers.

$$\begin{aligned} N_B Q|\psi> &= QN_B|\psi> + Q|\psi> = (n_B + 1)Q|\psi> \\ N_F Q|\psi> &= QN_F|\psi> - Q|\psi> = (n_F - 1)Q|\psi> \\ N_B Q^\dagger|\psi> &= Q^\dagger N_B|\psi> - Q^\dagger|\psi> = (n_B - 1)Q|\psi> \\ N_F Q^\dagger|\psi> &= Q^\dagger N_B|\psi> + Q^\dagger|\psi> = (n_F + 1)Q|\psi> \end{aligned} \tag{4.7}$$

The above relations mean that the operator Q acting on a fermion state increases the boson number by 1 and decreases the fermion number by 1; the effect of $Q^\dagger$ is the opposite. Defining

$$|n_B, n_F> = \frac{(b^\dagger)^{n_B}}{\sqrt{n_B!}}(a^\dagger)^{n_F}|0> \tag{4.8}$$

it is easy to verify that

$$Q|n_B, n_F> = \begin{cases} \sqrt{n_B + 1}|n_B + 1, n_F - 1> & n_F \neq 0 \\ 0 & n_F = 0 \end{cases} \tag{4.9a}$$

$$Q^\dagger|n_B, n_F> = \begin{cases} \sqrt{n_B}|n_B - 1, n_F + 1> & n_B \neq 0,\ n_F \neq 1 \\ 0 & n_B = 0 \text{ or } n_F = 1 \end{cases} \tag{4.9b}$$

Finally,

$$\{Q,\, Q^\dagger\} = \frac{1}{\omega} H \tag{4.10}$$

and we see that a set of operators ($Q,\ Q^\dagger,\ H$) constitute a closed algebra. Therefore, if the SUSY exists, its operation exchanges the boson and the fermion, and hence there are sets of multiplets whose members consist of both fermions and bosons with equal mass. Equation (4.10) shows, figuratively, that the operator Q is like a square root of the Hamiltonian, just as the Klein–Gordon equation was factorized into a product of two Dirac equations. As the Hamiltonian is the time translation operator, and Q changes the spin of the states, they are both external (i.e., spacetime) symmetries.

Another feature of the SUSY is a constraint that makes the vacuum energy vanish. As $a|0>= b|0>= 0$,

$$Q|0>= Q^\dagger|0>= 0 \tag{4.11}$$

If one takes vacuum expectation value (VEV) of Eq. (4.10), one obtains

$$<0|H|0>= 0 \tag{4.12}$$

Equation 4.12 is valid if H is the hamiltonian of the field theory. The reason for vanishing vacuum energy is that the zero-point energy created by the boson is exactly canceled by the zero-point energy of the fermion.

The zero-point energy, which is inherent to the quantum field theory, is generally ignored for the reason that only differences from the reference point are observable. However, we learned that the zero-point energy is real as demonstrated by the Casimir effect (see Section 5.5 of [1]). Bosons produce a divergent vacuum energy when summed over all degrees of freedom of harmonic oscillators. Fermions also produce infinite vacuum energy but with the opposite sign. The SUSY provides a solid foundation for the vanishing of the vacuum energy.

4.1.2
Field Theoretical Operators

Now, let us consider a relativistic version of the toy model, i.e., the SUSY operator in the field theory. As is clear from the toy model, to close the operator algebra, one needs both commutators and anti-commutators. Lie algebra, which is a mathematical framework for discussing the group structure of field operators, has to be extended to include the anti-commutation relations. One also has to extend the notion of conventional numbers to include anti-commuting numbers known as the *Grassmann numbers*. They are necessary ingredients to complete the algebra in a self-consistent mathematical framework. Here, we first introduce commutation relations among the SUSY operators that appear in the field theory without proof and discuss what they mean. The proof will be given later.

Let us, for simplicity, consider a SUSY operator Q_α having spin 1/2 and only one internal degree of freedom. If we treat it as the four-component spinor, it is a

Majorana spinor that satisfies the following commutation relations:

$$[Q_\alpha, M^{\mu\nu}] = \frac{i}{2}\left(\sigma^{\mu\nu}Q\right)_\alpha \tag{4.13a}$$

$$[Q_\alpha, P^\mu] = 0 \tag{4.13b}$$

$$\{Q_\alpha, \overline{Q}_\beta\} = 2(\gamma^\mu)_{\alpha\beta}P_\mu \tag{4.13c}$$

$$\{Q_\alpha, Q_\beta\} = \{\overline{Q}_\alpha, \overline{Q}_\beta\} = 0 \tag{4.13d}$$

where $\sigma^{\mu\nu} = (1/2i)[\gamma^\mu, \gamma^\nu]$, $\overline{Q} = Q^\dagger\gamma^0$ and we omitted the index of the internal degree of freedom. The first line shows that the Q_α transforms as a spinor, and the second shows this is a conserved operator. It also means that Q_α commutes with $P_\mu P^\mu = m^2$, that is, operation of Q_α does not change mass of the field. Fermions and bosons that belong to the same multiplet must have the same mass.

Let us call a multiplet including both fermions and bosons as a supermultiplet. and corresponding fields as superfields. Each supermultiplet contains an equal number of fermion and boson degrees of freedom. To prove this, one considers an operator $P_s \equiv (-1)^{2s}$, where s is the spin angular momentum. The operator P_s has eigenvalue +1 for the bosonic state and −1 for the fermionic state. Consider action of Q or $Q^\dagger$ on a state having the energy–momentum P_μ. It changes a fermionic state to a bosonic and vice versa within a supermultiplet. Therefore, the operator P_s anticommutes with the SUSY operator Q and $Q^\dagger$. One also has a completeness relation $\sum_i |i\rangle\langle i| = 1$, with i running through all the states within a given supermultiplet. Now consider an operator $(-1)^{2s}P_\mu$ whose trace is $p_\mu(n_B - n_F)$ because

$$\sum_i \langle i|(-1)^{2s}P_\mu|i\rangle = p_\mu \sum_i \langle i|(-1)^{2s}|i\rangle = p_\mu(n_B - n_F) \tag{4.14a}$$

On the other hand,

$$\begin{aligned}\sum_i \langle i|(-1)^{2s}P_\mu|i\rangle &\sim \sum_i \langle i|(-1)^{2s}Q\overline{Q}|i\rangle + \sum_i \langle i|(-1)^{2s}\overline{Q}Q|i\rangle\\ &= \sum_i \langle i|(-1)^{2s}Q\overline{Q}|i\rangle + \sum_i\sum_j \langle i|(-1)^{2s}\overline{Q}|j\rangle\langle j|Q|i\rangle\\ &= \sum_i \langle i|(-1)^{2s}Q\overline{Q}|i\rangle + \sum_j \langle j|Q(-1)^{2s}\overline{Q}|j\rangle\\ &= \sum_i \langle i|(-1)^{2s}Q\overline{Q}|i\rangle - \sum_j \langle j|(-1)^{2s}Q\overline{Q}|j\rangle\\ &= 0\end{aligned} \tag{4.14b}$$

where we suppressed the gamma matrix. The first line uses Eq. (4.13c), the second and the third lines the completeness condition, and the fourth line the anticommutativity of the operator $(-1)^{2s}$ with Q. Therefore, we have $n_B = n_F$ in any given supermultiplet.

Eq. (4.13c) means that a product $Q\overline{Q}$ is equivalent to the energy–momentum operator which is the generator of the spacetime translation. Coordinate transformations (translation and rotation) such as Lorentz transformation are global, but when localized, their transformation matrices become functions of spacetime, which means that the transformation becomes a general coordinate transformation including acceleration. According to the equivalence principle of general relativity, gravity is nothing but an inertial force generated by transferring an observer from an inertial frame to an accelerating frame. As the SUSY algebra includes the energy–momentum operator, it follows that a gauged (i.e., localized) SUSY algebra includes gravity. Therefore, the gauge theory based on the SUSY, which is referred to as *supergravity* (*SUGRA*), is a candidate of unified theories of all the forces.

4.2
Two-Component Formalism

4.2.1
Majorana Fields

In handling the SUSY algebra,[1] it is often more convenient to use two-component fermion fields $\left(\phi^T = (\phi_1, \phi_2)\right)$ instead of conventional four-component Dirac spinors. In this expression, the Dirac field is equivalent to a set of two Weyl fields with equal mass[2] just as a complex scalar field is a set of two real scalars. To construct the two-component formalism, we begin with the four-component Weyl representation for the γ matrices.

$$\gamma^\mu = \begin{bmatrix} 0 & \sigma^\mu \\ \overline{\sigma}^\mu & 0 \end{bmatrix}, \qquad \gamma^5 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \tag{4.15}$$

$$\sigma^\mu \equiv (1, \boldsymbol{\sigma}), \quad \overline{\sigma}^\mu = (1, -\boldsymbol{\sigma}) \tag{4.16}$$

$$\phi^{(4)} = \frac{1-\gamma^5}{2}\psi^{(4)} \equiv P_L\psi^{(4)} = \begin{bmatrix} \phi \\ 0 \end{bmatrix}, \qquad \chi^{(4)} = \frac{1+\gamma^5}{2}\psi^{(4)} \equiv P_R\psi^{(4)} = \begin{bmatrix} 0 \\ \chi \end{bmatrix} \tag{4.17}$$

where the index $^{(4)}$ is attached for the four-component spinor. ϕ and χ represent chirality $\mp$ or left- and right-handed particles, respectively, which transform as (1/2, 0) and (0, 1/2) in the $SU(2) \times SU(2)$ representation of the Lorentz group (see Appendix A of [1]).

The four-component Dirac equation can be expressed in terms of the two-component spinors ϕ, χ as

$$\psi^{(4)} = \begin{bmatrix} \phi \\ \chi \end{bmatrix}, \qquad \overline{\psi}^{(4)} = \psi^\dagger\gamma^0 = (\chi^\dagger, \phi^\dagger) \tag{4.18}$$

$$\begin{aligned} \mathscr{L}_{\text{Dirac}} &= \overline{\psi}^{(4)}(i\gamma^\mu\partial_\mu - m)\psi^{(4)} \\ &= (\chi^\dagger, \phi^\dagger)\begin{bmatrix} -m & i(\partial_0 + \boldsymbol{\sigma}\cdot\nabla) \\ i(\partial_0 - \boldsymbol{\sigma}\cdot\nabla) & -m \end{bmatrix}\begin{bmatrix} \phi \\ \chi \end{bmatrix} \\ &= \phi^\dagger\overline{\sigma}^\mu i\partial_\mu\phi + \chi^\dagger\sigma^\mu i\partial_\mu\chi - m(\phi^\dagger\chi + \chi^\dagger\phi) \end{aligned} \tag{4.19}$$

which shows that $\phi^\dagger\sigma^\mu\phi$ and $\chi^\dagger\sigma^\mu\chi$ are Lorentz vectors. In the massless limit, the Dirac equation reduces to $\overline{\sigma}^\mu i\partial_\mu\phi = 0$ and $\sigma^\mu i\partial_\mu\chi = 0$. It means that $\boldsymbol{\sigma}\cdot\mathbf{p}\phi = -E\phi$ and $\boldsymbol{\sigma}\cdot\mathbf{p}\chi = E\chi$ which confirms that $\phi(\chi)$ is left(right)-handed. If one imposes the Majorana condition

---

[1] For introductory or review articles and books, the reader may refer to [363, 405–408].

[2] Strictly, the word "Weyl spinor" applies only to massless field. However, we use it loosely to mean two-component spinors.

$$\psi^{(4)} = -(\psi^{(4)})^c = -C\overline{\psi}^{(4)\,T},^{3)} \qquad C = \begin{bmatrix} i\sigma_2 & 0 \\ 0 & -i\sigma_2 \end{bmatrix} \tag{4.20}$$

then, we can write $\phi = -i\sigma_2\chi^*,\ \chi = i\sigma_2\phi^*$.

The Lorentz transformation of the two-component spinor is given by the 2×2 complex matrix M with $\det[M] = 1$ which forms a group referred to as $SL(2, C)$ (see Appendix A of [1]).

$$\phi \overset{L}{\rightarrow} \phi' = \begin{bmatrix} \phi'_1 \\ \phi'_2 \end{bmatrix} = \begin{bmatrix} M_1^1 & M_1^2 \\ M_2^1 & M_2^2 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} \equiv M\phi \tag{4.21a}$$

$$\det[M] = M_1^1 M_2^2 - M_1^2 M_2^1 = 1 \quad \rightarrow \quad \varepsilon^{\alpha\beta} M_\alpha^\gamma M_\beta^\delta = \varepsilon^{\gamma\delta} \tag{4.21b}$$

where

$$\varepsilon^{12} = -\varepsilon^{21} = -\varepsilon_{12} = \varepsilon_{21} = 1, \quad \varepsilon^{\alpha\alpha} = \varepsilon_{\alpha\alpha} = 0 \quad (\alpha = 1,\, 2) \tag{4.21c}$$

The complex conjugate field ϕ^* transforms as $\phi^{*'} = M^*\phi^*$. M^* is not related to M by SMS^{-1} for any matrix S and hence is mathematically inequivalent. The complex conjugate of the left-handed spinor $\phi^* = -i\sigma_2\chi$ is a right-handed Weyl spinor. We therefore denote indices of the complex field ϕ^* with dots $(\phi_\alpha)^* \equiv (\overline{\phi})_{\dot{a}}$ to distinguish them from undotted indices used for ϕ_α.

It is convenient to treat $\varepsilon^{\alpha\beta}, \varepsilon_{\alpha\beta}$ as metric tensors conventionally used in the four-component relativistic equation. It can raise (or lower) the lower (or upper) indices and sum is to be understood when the same upper and lower indices appear at the same time.

$$\begin{gathered} \varepsilon_{\alpha\beta}\varepsilon^{\beta\gamma} \equiv \textstyle\sum_\beta \varepsilon_{\alpha\beta}\varepsilon^{\beta\gamma} = \delta_\alpha{}^\gamma \\ \phi^\alpha \equiv \varepsilon^{\alpha\beta}\phi_\beta = \textstyle\sum_\beta \varepsilon^{\alpha\beta}\phi_\beta, \quad \phi_\alpha \equiv \varepsilon_{\alpha\beta}\phi^\beta = \textstyle\sum_\beta \varepsilon_{\alpha\beta}\phi^\beta \end{gathered} \tag{4.22}$$

Noting that $(i\sigma_2)^{\alpha\beta} = \varepsilon^{\alpha\beta}$, $(-i\sigma_2)_{\alpha\beta} = \varepsilon_{\alpha\beta}$, we define

$$(\phi_\alpha)^* \equiv \overline{\phi}_{\dot{\alpha}} \quad \rightarrow \quad (i\sigma_2\phi^*)^{\dot{\alpha}} = \varepsilon^{\dot{\alpha}\dot{\beta}}\,\overline{\phi}_{\dot{\beta}} = \overline{\phi}^{\dot{\alpha}} \tag{4.23}$$

Thus, the four-component Majorana field can be expressed as

$$\psi^{(4)} = \begin{bmatrix} \phi_\alpha \\ \overline{\phi}^{\dot{\alpha}} \end{bmatrix} \tag{4.24}$$

The transformation matrix for $\overline{\phi}^{\dot{\alpha}}$ is given by

$$\overline{\phi}^{\dot{\alpha}} \overset{L}{\rightarrow} \overline{\phi}^{\dot{\alpha}'} = N^{\dot{\alpha}}{}_{\dot{\beta}}\,\overline{\phi}^{\dot{\beta}}, \qquad N = (i\sigma_2)M^*(-i\sigma_2) \tag{4.25}$$

In our representation, the Dirac spinor is written in terms of two two-component, complex, anti-commuting variables (i.e., Grassmann numbers: see boxed paragraph) ϕ_α and $(\chi^\dagger)^{\dot{\alpha}}$ with two distinct types of spinor indices.

---

[3] The minus sign is adopted here so that the C matrix representation in Eq. (4.20) and our definition $\chi = i\sigma_2\phi^*$ are consistent.

Arithmetics of Grassmann numbers: (see Section 11.4.1 of [1])
Two independent Grassmann numbers are anti-commuting, that is, $\eta_i\eta_j = -\eta_j\eta_i$. In particular, $\eta^2 = 0$, which means that the Grassmann numbers cannot have length or an absolute value.
Taylor expansion: Any function of Grassmann numbers including ordinary c-numbers (denoted in Latin characters) has a simple Taylor expansion:

$$\begin{aligned} f(\eta) &= a + b\eta \\ f(\eta_1, \eta_2) &= a + b_1\eta_1 + b_2\eta_2 + c\eta_1\eta_2 \end{aligned} \tag{4.26}$$

In particular, an exponential function can be written as

$$e^{a\eta} = 1 + a\eta \tag{4.27}$$

Complex conjugate: For any two real Grassmann numbers (ξ, η),

$$\zeta \equiv \xi + i\eta \quad \to \quad (\zeta_1\zeta_2)^* = \zeta_2^*\zeta_1^* \tag{4.28}$$

Derivative :

$$\begin{aligned} \text{Left derivative: } \frac{\partial}{\partial\eta_i}(\eta_j\eta_k) &= \delta_{ij}\eta_k - \delta_{ik}\eta_j \\ \text{Right derivative: } \frac{\partial}{\partial\eta_i}(\eta_j\eta_k) &= \delta_{ik}\eta_j - \delta_{ij}\eta_k \end{aligned} \tag{4.29}$$

The left derivatives are used throughout in this book.

$$\frac{\partial}{\partial\eta_i}\frac{\partial}{\partial\eta_j} = -\frac{\partial}{\partial\eta_j}\frac{\partial}{\partial\eta_i} \quad \to \quad \left(\frac{\partial}{\partial\eta}\right)^2 = 0 \tag{4.30}$$

Integral is equivalent to derivative:

$$\begin{aligned} &\int d\eta_i = 0, \quad \int d\eta_i\eta_j = \delta_{ij}, \quad \int d\eta_i\eta_j\eta_k = \delta_{ij}\eta_k - \delta_{ik}\eta_j \\ &\int d\eta_i f(\eta_j) = \int d\eta_i(a + b\eta_j) = b\delta_{ij} = \frac{\partial}{\partial\eta_i}f(\eta_j) \\ &\int d\eta_1 d\eta_2 f(\eta_1, \eta_2) = \int d\eta_1 d\eta_2(a + b_1\eta_1 + b_2\eta_2 + c\eta_1\eta_2) \\ &\quad = -c = \frac{d^2}{d\eta_1 d\eta_2}f(\eta_1, \eta_2) \end{aligned} \tag{4.31}$$

Jacobian of the Grassmann numbers is inverse of c-number Jacobian:

$$\int d\eta f(\eta) = \frac{\partial}{\partial\eta}f(\eta) = a\frac{\partial f(\eta'/a)}{\partial\eta'} = a\int d\eta' f\left(\frac{\eta'}{a}\right) \tag{4.32}$$

For $\xi_i = \sum_j U_{ij}\eta_j$ where U_{ij} are c-numbers:

$$\begin{aligned} &\int \prod_{i=1}^{N} d\eta_i f(\eta_1, \ldots, \eta_N) = J_G \int \prod_{i=1}^{N} d\xi_i g(\xi_1, \ldots, \xi_N) \\ &g(\xi) = f(U^{-1}\xi), \quad J_G = \frac{\partial(\xi_1, \ldots, \xi_N)}{\partial(\eta_1, \ldots, \eta_N)} = \det[U_{ij}] \end{aligned} \tag{4.33}$$

δ **function**:

$$\delta(\eta) = \eta, \quad \delta(f(\eta)) = \frac{\partial f(\eta)}{\partial\eta}\delta(\eta) \tag{4.34}$$

In particular, scalar products of any two Weyl spinors are defined by

$$\begin{aligned}
&\chi\phi \equiv \varepsilon^{\alpha\beta}\chi_\beta\phi_\alpha = \chi^\alpha\phi_\alpha = -\chi_\alpha\phi^\alpha = \phi^\alpha\chi_\alpha = \phi\chi\\
&\bar{\chi}\bar{\phi} \equiv \varepsilon_{\dot{\alpha}\dot{\beta}}\bar{\chi}^{\dot{\beta}}\bar{\phi}^{\dot{\alpha}} = \varepsilon^{\dot{\alpha}\dot{\beta}}\bar{\chi}_{\dot{\alpha}}\bar{\phi}_{\dot{\beta}} = \bar{\chi}_{\dot{\alpha}}\bar{\phi}^{\dot{\alpha}} = -\bar{\chi}^{\dot{\alpha}}\bar{\phi}_{\dot{\alpha}} = \bar{\phi}\bar{\chi}\\
&\bar{\chi}\bar{\phi} = \chi^*_{\dot{1}}\phi^{\dot{1}*} + \chi_{\dot{2}}\phi^{\dot{2}*} = (\phi^1\chi_1 + \phi^2\chi_2)^* = (\chi\phi)^*\\
&\chi\phi = (\bar{\chi}\bar{\phi})^*
\end{aligned} \tag{4.35}$$

and are Lorentz invariant. Notice that summation convention is different for dotted variables and that $\chi_\alpha\phi^\alpha = -\chi^\alpha\phi_\alpha \neq \chi\phi$. The ordering is important for the anti-commuting fields. For $\chi = \phi = \theta$, one has

$$\begin{aligned}
&\theta^\alpha\theta^\beta = -\tfrac{1}{2}\varepsilon^{\alpha\beta}\theta^\gamma\theta_\gamma = -\tfrac{1}{2}\varepsilon^{\alpha\beta}\theta\theta, \qquad \theta_\alpha\theta_\beta = \tfrac{1}{2}\varepsilon_{\alpha\beta}\theta\theta\\
&\bar{\theta}^{\dot{\alpha}}\bar{\theta}^{\dot{\beta}} = \tfrac{1}{2}\varepsilon^{\dot{\alpha}\dot{\beta}}\overline{\theta\theta} = -\tfrac{1}{2}\varepsilon_{\dot{\alpha}\dot{\beta}}\bar{\theta}_{\dot{\gamma}}\bar{\theta}^{\dot{\gamma}}, \qquad \bar{\theta}_{\dot{\alpha}}\bar{\theta}_{\dot{\beta}} = -\tfrac{1}{2}\varepsilon_{\dot{\alpha}\dot{\beta}}\bar{\theta}\bar{\theta}
\end{aligned} \tag{4.36}$$

Notice that our $\theta\theta$ is defined as a scalar product of two Grassmann spinors and is not to be confused with the square of ordinary Grassmann numbers for which $\theta^2 = 0$.

Putting $\chi_\alpha = \bar{\phi}^{\dot{\alpha}}$, $\chi^\dagger_{\dot{\alpha}} = \phi^\alpha$ in Eq. (4.19), the Majorana Lagrangian can be written as

$$\mathscr{L}_{\text{Majorana}} = \frac{1}{2}\left[\bar{\phi}_{\dot{\alpha}}\bar{\sigma}^{\mu\,\dot{\alpha}\beta}i\partial_\mu\phi_\beta + \phi^\alpha\sigma^\mu_{\alpha\dot{\beta}}\,i\partial_\mu\bar{\phi}^{\dot{\beta}} - m(\bar{\phi}_{\dot{\alpha}}\bar{\phi}^{\dot{\alpha}} + \phi^\alpha\phi_\alpha)\right] \tag{4.37}$$

The equations of motion for Majorana fields are as follows:

$$\bar{\sigma}^{\mu\,\dot{\alpha}\beta}i\partial_\mu\phi_\beta - m\bar{\phi}^{\dot{\alpha}} = 0 \tag{4.38a}$$

$$\sigma^\mu_{\alpha\dot{\beta}}i\partial_\mu\bar{\phi}^{\dot{\beta}} - m\phi_\alpha = 0 \tag{4.38b}$$

The above equations show that $\sigma^\mu_{\alpha\dot{\beta}}$ has natural lower indices, whereas $\bar{\sigma}^{\mu\,\dot{\alpha}\beta}$ has natural upper indices. One also sees that $\sigma^\mu_{\alpha\dot{\alpha}}$ convert a dotted spinor to an undotted spinor and $\bar{\sigma}^{\mu\,\dot{\alpha}\alpha}$ works in the opposite direction. That is, the Pauli matrices $\bar{\sigma}^\mu$ convert spinors belonging to (1/2, 0) to those in (0, 1/2) and vice versa for σ^μ.

Thus, both $\bar{\phi}\bar{\sigma}^\mu\phi$ and $\phi\sigma^\mu\bar{\phi}$ are Lorentz contravariant vectors. Similarly, $\phi\sigma^{\mu\nu}\chi$ and $\bar{\phi}\bar{\sigma}^{\mu\nu}\bar{\chi}$ are Lorentz tensors where

$$\begin{aligned}
&\Sigma^{\mu\nu} = \tfrac{i}{2}[\gamma^\mu, \gamma^\nu] = \begin{bmatrix}\sigma^{\mu\nu} & 0\\ 0 & \bar{\sigma}^{\mu\nu}\end{bmatrix}\\
&\sigma^{\mu\nu} = \tfrac{i}{2}(\sigma^\mu\bar{\sigma}^\nu - \sigma^\nu\bar{\sigma}^\mu), \qquad \bar{\sigma}^{\mu\nu} = \tfrac{i}{2}(\bar{\sigma}^\mu\sigma^\nu - \bar{\sigma}^\nu\sigma^\mu)
\end{aligned} \tag{4.39}$$

Conversion of the four-component spinor to the two-component spinor : Without loss of generality, the four-component Dirac spinor can be redefined as

$$\psi_D^{(4)} = \begin{bmatrix}\phi_\alpha\\ \bar{\chi}^{\dot{\alpha}}\end{bmatrix} \tag{4.40}$$

where both ϕ and χ are left-handed spinors. Thus, the four component Dirac spinor is expressed in terms of two left-handed two-component spinors. In this representation, the Majorana condition is simply $\chi = \phi$. Some examples of four-component spinor representation expressed in the two-component representation are as follows:

$$\begin{aligned}
\overline{\psi}_i^{(4)}\psi_j^{(4)} &= \overline{\psi}_i^{(4)}P_L\psi_j^{(4)} + \overline{\psi}_i^{(4)}P_R\psi_j^{(4)} = \chi_i\phi_j + \overline{\phi}_i\overline{\chi}_j \\
\overline{\psi}_i^{(4)}\gamma^\mu\psi_j^{(4)} &= \qquad\qquad\qquad\qquad\qquad\quad = \overline{\phi}_i\overline{\sigma}^\mu\phi_j + \chi_i\sigma^\mu\overline{\chi}_j \\
\overline{\psi}_i^{(4)}\Sigma^{\mu\nu}\psi_j^{(4)} &= \qquad\qquad\qquad\qquad\qquad\quad = \chi_i\sigma^{\mu\nu}\phi_j + \overline{\phi}_i\overline{\sigma}^{\mu\nu}\overline{\chi}_j
\end{aligned} \tag{4.41}$$

In the above expressions, $\phi_i\chi_j$, and so on, are scalars made of Weyl spinors in the same sense as those in Eq. (4.35), and indices i, j are not spinor indexes but indicators just to distinguish different spinors.

Useful relations:

$$\varepsilon^{\alpha\beta}\frac{\partial}{\partial\theta^\beta} = -\frac{\partial}{\partial\theta_\alpha}, \qquad \frac{\partial}{\partial\theta^\alpha} = -\varepsilon_{\alpha\beta}\frac{\partial}{\partial\theta_\beta} \tag{4.42}$$

Proof:

$$\varepsilon^{\alpha\beta}\frac{\partial\theta^\gamma}{\partial\theta^\beta} = \varepsilon^{\alpha\beta}\delta_\beta^\gamma = \varepsilon^{\alpha\gamma}$$

$$-\frac{\partial\theta^\gamma}{\partial\theta_\alpha} = -\frac{\partial\varepsilon^{\gamma\delta}\theta_\delta}{\partial\theta_\alpha} = -\varepsilon^{\gamma\delta}\delta_\delta^\alpha = -\varepsilon^{\gamma\alpha} = \varepsilon^{\alpha\gamma}$$

Similarly

$$\varepsilon^{\dot\alpha\dot\beta}\frac{\partial}{\partial\overline{\theta}^{\dot\beta}} = -\frac{\partial}{\partial\overline{\theta}_{\dot\alpha}}, \qquad \frac{\partial}{\partial\overline{\theta}^{\dot\alpha}} = -\varepsilon_{\dot\alpha\dot\beta}\frac{\partial}{\partial\overline{\theta}_\beta} \tag{4.43}$$

Relations between σ^μ and $\overline{\sigma}^\mu$:

$$\sigma^{\mu\alpha\dot\delta} = \varepsilon^{\alpha\beta}\sigma^\mu_{\beta\dot\gamma}\varepsilon^{\dot\delta\dot\gamma} = -\varepsilon^{\alpha\beta}\sigma^\mu_{\beta\dot\gamma}\varepsilon^{\dot\gamma\dot\delta} = -[(i\sigma_2)\sigma^\mu(i\sigma_2)]^{\alpha\dot\delta} = [\overline{\sigma}^{\mu T}]^{\alpha\dot\delta} = \overline{\sigma}^{\mu\dot\delta\alpha} \tag{4.44a}$$

Similarly

$$\overline{\sigma}^\mu_{\dot\alpha\delta} = -\varepsilon_{\dot\alpha\dot\beta}\overline{\sigma}^{\mu\dot\beta\gamma}\varepsilon_{\gamma\delta} = \sigma^\mu_{\delta\dot\alpha} \tag{4.44b}$$

$$\begin{aligned}
&(\sigma^\mu)_{\alpha\dot\beta}(\overline{\sigma}^\nu)^{\dot\beta\beta} + (\sigma^\nu)_{\alpha\dot\beta}(\overline{\sigma}^\mu)^{\dot\beta\beta} \equiv (A^{\mu\nu})_\alpha{}^\beta = 2g^{\mu\nu}\delta_\alpha{}^\beta \\
&(\overline{\sigma}^\mu)^{\dot\alpha\beta}(\sigma^\nu)_{\beta\dot\beta} + (\overline{\sigma}^\nu)^{\dot\alpha\beta}(\sigma^\mu)_{\beta\dot\beta} \equiv (B^{\mu\nu})^{\dot\alpha}{}_{\dot\beta} = 2g^{\mu\nu}\delta^{\dot\alpha}{}_{\dot\beta} \\
&(A^{\mu\nu})^{\alpha\beta} = \ \ 2g^{\mu\nu}\varepsilon^{\alpha\beta}, \quad (A^{\mu\nu})_{\alpha\beta} = -2g^{\mu\nu}\varepsilon_{\alpha\beta} \\
&(B^{\mu\nu})^{\dot\alpha\dot\beta} = -2g^{\mu\nu}\varepsilon^{\dot\alpha\dot\beta}, \quad (B^{\mu\nu})_{\dot\alpha\dot\beta} = \ \ 2g^{\mu\nu}\varepsilon_{\dot\alpha\dot\beta}
\end{aligned} \tag{4.45}$$

Completeness relation: Any 2 × 2 matrix A can be expanded in terms of Pauli matrices.

$$A = \frac{1}{2}\Big(\mathrm{Tr}[A]\mathbf{1} + \mathrm{Tr}[\sigma_i A]\sigma_i\Big) \tag{4.46}$$

We will rewrite the above expression using dotted and undotted indices. An expression sandwiched by two Grassmann numbers $\theta A\overline{\theta} = \theta^\alpha A_{\alpha\dot\beta}\overline{\theta}^{\dot\beta}$ frequently appears in the following. Matching the dotted and undotted indices, Eq. (4.46) can be rewritten as

$$\begin{aligned} A_{\alpha\dot\beta} &= \tfrac{1}{2}\left(\mathrm{Tr}[A]\mathbf{1}_{\alpha\dot\beta} + \mathrm{Tr}[\sigma_i A](\sigma_i)_{\alpha\dot\beta}\right) \\ &= \tfrac{1}{2}\left((\mathbf{1})^{\dot\gamma\delta}A_{\delta\dot\gamma}\mathbf{1}_{\alpha\dot\beta} - (\sigma_i)^{\dot\gamma\delta}A_{\delta\dot\gamma}(-\sigma_{i\alpha\dot\beta})\right) \\ &= \tfrac{1}{2}\sigma_{\mu\alpha\dot\beta}(\overline{\sigma}^{\mu\dot\gamma\delta}A_{\delta\dot\gamma}) \end{aligned} \tag{4.47}$$

As A is an arbitrary matrix, it follows that

$$\sigma_{\mu\alpha\dot\beta}\overline{\sigma}^{\mu\dot\gamma\delta} = 2\delta_\alpha{}^\delta\delta_{\dot\beta}{}^{\dot\gamma} \tag{4.48}$$

Identities:

$$\nabla_\alpha\theta\theta = \frac{\partial\theta\theta}{\partial\theta^\alpha} = 2\theta_\alpha, \qquad \overline{\nabla}_{\dot\alpha}\overline{\theta\theta} = \frac{\partial\overline{\theta\theta}}{\partial\overline{\theta}^{\dot\alpha}} = -2\overline{\theta}_{\dot\alpha} \tag{4.49}$$

$$\begin{aligned} \because\quad \tfrac{\partial\theta\theta}{\partial\theta^\alpha} &= \tfrac{\partial\theta^\beta\theta_\beta}{\partial\theta^\alpha} = \delta_\alpha{}^\beta\theta_\beta - \theta^\beta\tfrac{\partial}{\partial\theta^\alpha}(\varepsilon_{\beta\gamma}\theta^\gamma) = \theta_\alpha - \theta^\beta\varepsilon_{\beta\gamma}\delta_\alpha{}^\gamma \\ &= \theta_\alpha - \theta^\beta\varepsilon_{\beta\alpha} = \theta_\alpha + \theta^\beta\varepsilon_{\alpha\beta} = \theta_\alpha + \theta_\alpha = 2\theta_\alpha \end{aligned} \tag{4.50}$$

$$\nabla^\alpha\nabla_\alpha\theta\theta = 4, \qquad \overline{\nabla}_{\dot\alpha}\overline{\nabla}^{\dot\alpha}\overline{\theta\theta} = 4 \tag{4.51}$$

$$(\theta\sigma^\mu\overline{\theta})(\theta\sigma^\nu\overline{\theta}) = \frac{1}{2}g^{\mu\nu}\theta\theta\overline{\theta\theta} \tag{4.52}$$

$$\begin{aligned} \because\quad (\theta\sigma^\mu\overline{\theta})(\theta\sigma^\nu\overline{\theta}) &= \theta^\alpha\sigma^\mu_{\alpha\dot\alpha}\overline{\theta}^{\dot\alpha}\theta^\beta\sigma^\nu_{\beta\dot\beta}\overline{\theta}^{\dot\beta} = -(\theta^\alpha\theta^\beta)(\overline{\theta}^{\dot\alpha}\overline{\theta}^{\dot\beta})\sigma^\mu_{\alpha\dot\alpha}\sigma^\nu_{\beta\dot\beta} \\ &= \tfrac{1}{4}\varepsilon^{\alpha\beta}\theta\theta\overline{\theta\theta}\varepsilon^{\dot\alpha\dot\beta}\sigma^\mu_{\alpha\dot\alpha}\sigma^\nu_{\beta\dot\beta} = \tfrac{1}{4}\theta\theta\overline{\theta\theta}\varepsilon^{\beta\alpha}\varepsilon^{\dot\beta\dot\alpha}\sigma^\mu_{\alpha\dot\alpha}\sigma^\nu_{\beta\dot\beta} = \tfrac{1}{4}\theta\theta\overline{\theta\theta}\sigma^{\mu\beta\dot\beta}\sigma^\nu_{\beta\dot\beta} \\ &= \tfrac{1}{4}\theta\theta\overline{\theta\theta}\overline{\sigma}^{\mu\dot\beta\beta}\sigma^\nu_{\beta\dot\beta} = \tfrac{1}{4}\theta\theta\overline{\theta\theta}(\overline{\sigma}^\mu\sigma^\nu)^{\dot\beta}{}_{\dot\beta} = \tfrac{1}{4}\theta\theta\overline{\theta\theta}\,\mathrm{Tr}[\overline{\sigma}^\mu\sigma^\nu] \\ &= \tfrac{1}{2}g^{\mu\nu}\theta\theta\overline{\theta\theta} \end{aligned} \tag{4.53}$$

4.2.2
SUSY Operators

The SUSY generators, by definition, transform a boson to a fermion or vice versa; hence, it is a fermionic operator and carries spin 1/2 (or half odd integer). [363,406,409–411]. The symmetry involves space rotation and must be a spacetime symmetry. There is a famous theorem that all the symmetry operators of the Lagrangian other than those of the Poincaré group that form a Lie algebra of commutators must be Lorentz scalars [412]. Generators of the Poincaré group satisfy the following commutation relations:

$$[M_{\mu\nu}, M_{\rho\sigma}] = i(g_{\nu\rho}M_{\mu\sigma} - g_{\mu\rho}M_{\nu\sigma} + g_{\mu\sigma}M_{\nu\rho} - g_{\nu\sigma}M_{\mu\rho}) \tag{4.54a}$$

$$[P_\mu, M_{\rho\sigma}] = i(g_{\mu\rho}P_\sigma - g_{\mu\sigma}P_\rho) \tag{4.54b}$$

$$[P_\mu, P_\nu] = 0 \tag{4.54c}$$

Therefore, if one wants a fermionic operator to be a part of some Lie algebra, which is the mathematical framework of the field operators, the Lie algebra must be generalized to contain anti-commutators. Such an extension of the Lie algebra is called a *graded* or *super algebra*. One needs Grassmann numbers to complete the algebra. To write down the extended part, let us, for simplicity, consider a two-component SUSY operator $Q_\alpha, \alpha = 1, 2$. Using the two-component representation, the commutator relations of the extended Lie algebra are defined by

$$[M^{\mu\nu}, Q_\alpha] = -\frac{1}{2}(\sigma^{\mu\nu})_{\alpha\beta}Q^\beta \tag{4.55a}$$

$$[M^{\mu\nu}, \overline{Q}^{\dot{\alpha}}] = -\frac{1}{2}(\overline{\sigma}^{\mu\nu})^{\dot{\alpha}\dot{\beta}}\overline{Q}_{\dot{\beta}} \tag{4.55b}$$

$$[Q_\alpha, P^\mu] = [\overline{Q}_{\dot{\alpha}}, P^\mu] = 0 \tag{4.55c}$$

$$\{Q_\alpha, \overline{Q}_{\dot{\beta}}\} = 2(\sigma^\mu)_{\alpha\dot{\beta}}P_\mu \tag{4.55d}$$

$$\{Q^\alpha, \overline{Q}^{\dot{\beta}}\} = 2(\overline{\sigma}^\mu)^{\dot{\beta}\alpha}P_\mu \tag{4.55e}$$

$$\{Q_\alpha, Q_\beta\} = \{\overline{Q}_{\dot{\alpha}}, \overline{Q}_{\dot{\beta}}\} = 0 \tag{4.55f}$$

Equations (4.55a)–(4.55f) constitute an algebra that closes within themselves and hence can serve as generators of the SUSY group.

The four-component version of Eqs. (4.55d) and (4.55e) is

$$\{Q^{(4)}, \overline{Q}^{(4)}\} = 2\gamma^\mu P_\mu \tag{4.56a}$$

$$Q^{(4)} = \begin{bmatrix} Q_\alpha \\ \overline{Q}^{\dot{\alpha}} \end{bmatrix}, \quad \overline{Q}^{(4)} = (Q^\alpha, \overline{Q}_{\dot{\alpha}}) \tag{4.56b}$$

where $\overline{Q}^{(4)} = Q^{(4)\dagger}\gamma^0$ and $Q_\alpha^\dagger = \overline{Q}_{\dot{\alpha}}$.

Justification of the above formulas comes from symmetry considerations [413]. The form of the right-hand side of Eqs. (4.55a) and (4.55b) can be determined considering the dimensionality and the Lorentz tensor structure with appropriate spinor legs. More directly, we note that the commutator of the angular momentum operator and any fermionic field ψ in four-component representations satisfies the commutation relation

$$[M^{\mu\nu}, \psi(y)] = \int d^3x\,[\psi^\dagger(x)S^{\mu\nu}\psi(x), \psi(y)] = -S^{\mu\nu}\psi(y) \tag{4.57a}$$

which becomes Eqs. (4.55a) and (4.55b) if we replace ψ with Q and rewrite it in the two-component representation $S^{\mu\nu} = (1/2)\sigma^{\mu\nu}$ for left-handed and $S^{\mu\nu} = (1/2)\overline{\sigma}^{\mu\nu}$ for right-handed spinors [see Eq. (4.39)].

The left-hand side of Eq. (4.55d) is a product of two spinors with dimension $E^{1/2}$. As $Q_\alpha, \overline{Q}^{\dot{\alpha}}$ transform like left- and right-handed spinors, that is, like $(1/2, 0)$ and $(0, 1/2)$, the product must transform like $(1/2, 1/2)$, that is, as a vector. Considering the spinor legs, the right-hand side must be proportional to $(\sigma^\mu_{\alpha\dot{\beta}}P_\mu)$. The factor 2 is a normalization constant.

The symmetry operator has the form $\sim\exp[i\overline{\theta}Q]$ or $\sim\exp[i\overline{Q}\theta]$. As Q_α, $\overline{Q}^\alpha$ has energy dimension $[E^{1/2}]$, it follows that $\overline{\theta}$ have dimension $[E^{-1/2}]$. As $Q, \overline{Q}$ are anti-commuting spinors, the parameters θ, $\overline{\theta}$ have to be also anti-commuting Grassmann numbers in the form of a Majorana spinor. For infinitesimal transformation ($\theta, \overline{\theta} = \delta\eta, \delta\overline{\eta}$) of a scalar field A and a spinor field ψ, we should have something like

$$\delta A \sim \delta\eta^{\alpha}\psi_{\alpha}, \qquad \delta\psi_{\alpha} \sim \delta\overline{\eta}^{\dot{\beta}}\overline{\sigma}^{\mu}_{\dot{\beta}\alpha} P_{\mu}A \tag{4.58}$$

The first term confirms that $\delta\eta$ has dimension $[E^{-1/2}]$ because the scalar and the fermionic field have dimension $[E]$, $[E^{3/2}]$, respectively. In the second term, the right-hand side has to be multiplied with P_{μ} to match the dimensionality, yet the whole expression has to be an undotted spinor, which leads to the form $\overline{\eta}^{\dot{\beta}}\overline{\sigma}^{\mu}_{\dot{\beta}\alpha} P_{\mu}$.

Equation (4.55c) can be proved as follows. The commutator of the Q with P_{μ} behaves like (1/2, 1), which has spin 3/2 or (1/2, 0) with dimension $\sim[E^{3/2}]$ but there are no spin 3/2 generators and we are left with

$$[Q_{\alpha}, P_{\mu}] = C(\sigma_{\mu})_{\alpha\dot{\beta}}\overline{Q}^{\dot{\beta}} \tag{4.59}$$

and its adjoint representation

$$[\overline{Q}^{\dot{\beta}}, P_{\mu}] = C^{*}(\overline{\sigma}_{\mu})^{\dot{\beta}\alpha}Q_{\alpha} \tag{4.60}$$

Notice that the equations of motion for the Majorana spinor Eqs. (4.38) exhibit a nice example of the above reasoning about the spinor legs. Making a double commutator with P_{μ}, P_{ν}, we obtain

$$[[Q_{\alpha}, P_{\mu}], P_{\nu}] = CC^{*}(\sigma_{\mu}\overline{\sigma}_{\nu})_{\alpha}{}^{\beta}Q_{\beta} \tag{4.61}$$

Substituting Eq. (4.61) in the Jacobi identity

$$[[Q_{\alpha}, P_{\mu}], P_{\nu}] + [[P_{\mu}, P_{\nu}], Q_{\alpha}] + [[P_{\nu}, Q_{\alpha}], P_{\mu}]] = 0 \tag{4.62}$$

and using $[P_{\mu}, P_{\nu}] = 0$, we obtain

$$CC^{*}(\sigma_{\mu}\overline{\sigma}_{\nu} - \sigma_{\nu}\overline{\sigma}_{\mu})_{\alpha}{}^{\beta}Q_{\beta} = 0 \tag{4.63}$$

This can be true only if $CC^{*} = 0$.

Finally, the requirement $[Q_{\alpha}, Q_{\beta}] = 0$ follows from vanishing of $[Q_{\alpha}, P_{\mu}]$ and the Jacobi identity for Q_{α}, Q_{β} and P_{μ}.

4.2.3
Superspace

The energy–momentum P_{μ} is a translation symmetry operator. It is defined by

$$\Phi(x+a) = e^{-ia^{\mu}P_{\mu}}\Phi(x)e^{ia^{\mu}P_{\mu}} \tag{4.64}$$

where $a = a^{\mu}$, ($\mu = 0-3$) is a set of four constant spacetime coordinates. In parallel to the energy–momentum operator, we consider the superfield operator Q_{α}, $\overline{Q}_{\dot{\alpha}}$ as generators of the translation in the fermionic coordinate system expressed by the Grassmann numbers. To match the four spacetime dimensions, we consider four fermionic coordinates as elements of a Majorana spinor or as a pair of two-component Weyl spinors. Points in superspace are identified by the coordinates

$$z = (x^{\mu},\ \theta^{\alpha},\ \overline{\theta}_{\dot{\alpha}}) = (x^{\mu}, \theta^{1}, \theta^{2}, \overline{\theta}_{\dot{1}}, \overline{\theta}_{\dot{2}}) \tag{4.65}$$

Commutators of the super-coordinates satisfy the following relations:

$$\begin{aligned} &\{\theta^{\alpha}, \theta^{\beta}\} = \{\overline{\theta}_{\dot{\alpha}}, \overline{\theta}_{\dot{\beta}}\} = \{\theta^{\alpha}, \overline{\theta}_{\dot{\beta}}\} = 0 \\ &[x^{\mu},\ \theta^{\alpha}] = [x^{\mu},\ \overline{\theta}_{\dot{\beta}}] = 0 \end{aligned} \tag{4.66}$$

The indices α and $\dot{\alpha}$ run from 1 to 2 and $(\theta^{\alpha})^{*} = \overline{\theta}^{\dot{\alpha}}$.

We define a translational operator (an element of the super group) $L(x, \theta, \overline{\theta})$. The superfield $\Phi(x, \theta^{\alpha}, \overline{\theta}_{\dot{\alpha}})$ is the operand of the SUSY generators, which is a function of spacetime and superspace coordinate variables.

$$\begin{aligned} &L(x, \theta, \overline{\theta}) \equiv \exp\left[i(-x^{\mu} P_{\mu} + \theta^{\alpha} Q_{\alpha} + \overline{\theta}_{\dot{\alpha}} \overline{Q}^{\dot{\alpha}})\right] \\ &\Phi(x, \theta, \overline{\theta}) = L(x, \theta, \overline{\theta})\, \Phi_0\, L^{-1}(x, \theta, \overline{\theta}) \end{aligned} \tag{4.67}$$

As a product of two group elements is also an element of the group, we have

$$\begin{aligned} &\exp\left[i(-x'^{\mu} P_{\mu} + \theta'^{\alpha} Q_{\alpha} + \overline{\theta}'_{\dot{\alpha}} \overline{Q}^{\dot{\alpha}})\right] \\ &\quad = \exp\left[i(-a^{\mu} P_{\mu} + \xi^{\alpha} Q_{\alpha} + \overline{\xi}_{\dot{\alpha}} \overline{Q}^{\dot{\alpha}})\right] \exp\left[i(-x^{\mu} P_{\mu} + \theta^{\alpha} Q_{\alpha} + \overline{\theta}_{\dot{\alpha}} \overline{Q}^{\dot{\alpha}})\right] \end{aligned} \tag{4.68}$$

Using the Baker–Campbell–Hausdorff formula

$$e^{A} e^{B} = e^{A+B+\left(\frac{1}{2}\right)[A,B]+\left(\frac{1}{12}\right)[A,[A,B]-\left(\frac{1}{12}\right)[B,[B,A]]+\cdots} \tag{4.69}$$

and the super Poincaré algebra Eqs. (4.55), we can calculate the r.h.s. of Eq. (4.68). It is easy to show that

$$\begin{aligned} &[\xi Q,\ \overline{\theta}\,\overline{Q}] = \ \ 2(\xi \sigma^{\mu} \overline{\theta}) P_{\mu} \\ &[\overline{\xi}\,\overline{Q},\ \theta Q] = -2(\theta \sigma^{\mu} \overline{\xi}) P_{\mu} \\ &[\xi Q,\ \theta Q] = [\overline{\xi}\,\overline{Q},\ \overline{\theta}\,\overline{Q}] = 0 \end{aligned} \tag{4.70}$$

The higher order terms do not appear. Equation (4.68) becomes

$$\begin{aligned} &\exp\left[i(-a^{\mu} P_{\mu} + \xi^{\alpha} Q_{\alpha} + \overline{\xi}_{\dot{\alpha}} \overline{Q}^{\dot{\alpha}})\right] \exp\left[i(-x^{\mu} P_{\mu} + \theta^{\alpha} Q_{\alpha} + \overline{\theta}_{\dot{\alpha}} \overline{Q}^{\dot{\alpha}})\right] \\ &\quad = \exp\left[-i(x^{\mu} + a^{\mu} + T^{\mu}) P_{\mu} + i(\theta^{\alpha} + \xi^{\alpha}) Q_{\alpha} + i(\overline{\theta}_{\dot{\alpha}} + \overline{\xi}_{\dot{\alpha}}) \overline{Q}^{\dot{\alpha}}\right] \\ &T^{\mu} = i(\theta \sigma^{\mu} \overline{\xi} - \xi \sigma^{\mu} \overline{\theta}) \end{aligned} \tag{4.71}$$

Thus, we have the transformation property of the supergenerator

$$\begin{aligned} S\Phi(x, \theta, \overline{\theta}) &\equiv L(0, \xi, \overline{\xi})\, \Phi(x, \theta, \overline{\theta})\, L^{-1}(0, \xi, \overline{\xi}) \\ &= \Phi\left(x^{\mu} + i(\theta \sigma^{\mu} \overline{\xi} - \xi \sigma^{\mu} \overline{\theta}), \theta^{\alpha} + \xi^{\alpha}, \overline{\theta}_{\dot{\alpha}} + \overline{\xi}_{\dot{\alpha}}\right) \end{aligned} \tag{4.72}$$

The important thing to notice is that translations in the supercoordinates necessarily induce translations in the spacetime coordinates. Therefore, interpretation of the supergenerator as a translation operator of the fermionic coordinates is valid, but unlike the spacetime translation, it is not simply expressed as $Q_\alpha \sim \partial/\partial\theta^\alpha$, $\overline{Q}^{\dot{\alpha}} \sim \partial/\partial\overline{\theta}_{\dot{\alpha}}$. For most purposes, it is sufficient to consider infinitesimal transformations. These can be written as

$$\begin{aligned}\delta_S(\xi,\overline{\xi})\Phi(x,\theta,\overline{\theta}) &= i\left[\xi Q + \overline{\xi}\overline{Q}, \Phi(x,\theta,\overline{\theta})\right]\\ &= \left[\xi^\alpha \frac{\partial}{\partial\theta^\alpha} + \overline{\xi}_{\dot{\alpha}} \frac{\partial}{\partial\overline{\theta}_{\dot{\alpha}}} - i\left\{(\xi\sigma^\mu\overline{\theta}) - (\theta\sigma^\mu\overline{\xi})\right\}\partial_\mu\right]\Phi(x,\theta,\overline{\theta})\end{aligned} \tag{4.73}$$

This corresponds to the following explicit representation of the SUSY generators [406, 409]:

$$iQ_\alpha \equiv \quad \nabla_\alpha - i\sigma^\mu_{\alpha\dot{\alpha}}\overline{\theta}^{\dot{\alpha}}\partial_\mu \tag{4.74a}$$

$$i\overline{Q}_{\dot{\alpha}} \equiv -\overline{\nabla}_{\dot{\alpha}} + i\theta^\alpha\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu \tag{4.74b}$$

where we introduced the nabla notation $\nabla_\alpha \equiv \partial/\partial\theta^\alpha$, $\overline{\nabla}_{\dot{\alpha}} \equiv \partial/\partial\overline{\theta}^{\dot{\alpha}} = -\varepsilon_{\dot{\alpha}\dot{\beta}}\partial/\partial\overline{\theta}_{\dot{\beta}}$ in order not to confuse with spacetime derivatives ∂_μ.

At this point, we remind the reader that a valid global symmetry should apply also to the derivatives of the field as both the field itself (denoted as f) and its derivatives appear in the Lagrangian. An important notion in the SUSY is that $\nabla_\alpha f$, $\overline{\nabla}_{\dot{\alpha}} f$ do not satisfy the global SUSY hence are not superfields. By global symmetry, we mean that SUSY operators acting on them satisfy the commutation relations defined by Eqs. (4.55). Notice that $\partial_\mu f$ is a superfield because it satisfies the global SUSY but $\nabla_\alpha f$, $\overline{\nabla}_{\dot{\alpha}} f$ are not. This can be seen by looking at its commutators with the symmetry operators.

$$[\nabla_\alpha, (\theta Q)] = Q_\alpha, \quad [\nabla_\alpha, (\overline{\theta}\overline{Q})] = \sigma^\mu_{\alpha\dot{\alpha}}\overline{\theta}^{\dot{\alpha}}\partial_\mu, \quad \ldots \tag{4.75a}$$

We have to find differential operators that commute with the SUSY operators. This can be achieved by defining the covariant derivatives

$$D_\alpha = \quad \nabla_\alpha + i\sigma^\mu_{\alpha\dot{\alpha}}\overline{\theta}^{\dot{\alpha}}\partial_\mu \tag{4.76a}$$

$$\overline{D}_{\dot{\alpha}} = -\overline{\nabla}_{\dot{\alpha}} - i\theta^\alpha\sigma^\mu_{\alpha\dot{\alpha}}\partial_\mu \tag{4.76b}$$

The covariant derivatives satisfy the following relations:

$$\{D_\beta, Q_\alpha\} = \{D_\beta, \overline{Q}_{\dot{\alpha}}\} = \{\overline{D}_{\dot{\beta}}, Q_\alpha\} = \{\overline{D}_{\dot{\beta}}, \overline{Q}_{\dot{\alpha}}\} = 0 \tag{4.77a}$$

$$\{D_\alpha, D_\beta\} = \{\overline{D}_{\dot{\alpha}}, \overline{D}_{\dot{\beta}}\} = 0 \tag{4.77b}$$

$$(D^3)^\alpha \equiv (D^\beta D_\beta)D^\alpha = 0, \qquad (\overline{D}^3)_{\dot{\alpha}} \equiv (\overline{D}_{\dot{\beta}}\overline{D}^{\dot{\beta}})\overline{D}_{\dot{\alpha}} = 0 \tag{4.77c}$$

$$\{D_\alpha, \overline{D}_{\dot{\beta}}\} = -2i\sigma^\mu_{\alpha\dot{\beta}}\partial_\mu \tag{4.77d}$$

Problem 4.1 *Prove Eqs. (4.77)*

For reference, we show that in the four-component representation, we have

$$\Theta^{(4)} \equiv \begin{bmatrix} \theta_\alpha \\ \overline{\theta}^{\dot\alpha} \end{bmatrix}, \qquad \overline{\Theta}^{(4)} \equiv (\theta^\alpha, \overline{\theta}_{\dot\alpha}) \tag{4.78}$$

$$\rightarrow \quad \left(\frac{\partial}{\partial\Theta}\right)^{(4)}_\alpha \equiv \left(\frac{\partial}{\partial\theta_\alpha}, \frac{\partial}{\partial\overline{\theta}^{\dot\alpha}}\right), \qquad \left(\frac{\partial}{\partial\overline{\Theta}}\right)^{(4)}_\alpha \equiv \begin{bmatrix} \dfrac{\partial}{\partial\theta^\alpha} \\ \dfrac{\partial}{\partial\overline{\theta}_{\dot\alpha}} \end{bmatrix} \tag{4.79}$$

We have

$$Q^{(4)} = \begin{bmatrix} Q_\alpha \\ \overline{Q}^{\dot\alpha} \end{bmatrix} = \begin{bmatrix} \nabla_\alpha \\ \overline{\nabla}^{\dot\alpha} \end{bmatrix} - i \begin{bmatrix} 0 & \sigma^{\mu}{}_{\alpha\dot\beta} \\ \overline{\sigma}^{\mu\,\dot\alpha\beta} & 0 \end{bmatrix} \begin{bmatrix} \theta_\beta \\ \overline{\theta}^{\dot\beta} \end{bmatrix} \partial_\mu = \left(\frac{\partial}{\partial\overline{\Theta}}\right)^{(4)} - i\gamma^{\mu}\Theta\partial_\mu \tag{4.80}$$

$$D^{(4)}_\alpha \equiv \begin{bmatrix} D_\alpha \\ \overline{D}^{\dot\alpha} \end{bmatrix} = \left(\frac{\partial}{\partial\overline{\Theta}}\right)^{(4)}_\alpha + i(\gamma^{\mu}\Theta)_\alpha\partial_\mu \tag{4.81a}$$

4.3
Chiral Superfield

Now, we consider a superfield $F(x, \theta, \overline{\theta})$, which is a function of x^μ, θ^α, $\overline{\theta}_{\dot\alpha}$. F could belong to an arbitrary representation of the Lorentz group but let us first consider a scalar field. The Taylor expansion in powers of θ and $\overline{\theta}$ terminates at finite order $\theta\theta, \overline{\theta\theta}$ as the variables are anti-commuting Grassmann numbers.

It can be expanded as [411]

$$\begin{aligned} F(x^\mu, \theta^\alpha, \overline{\theta}_{\dot\alpha}) = {} & A(x) + \theta\psi(x) + \overline{\theta\chi}(x) + \theta\theta m(x) + \overline{\theta\theta}n(x) + \theta\sigma^{\mu}\overline{\theta}V_\mu(x) \\ & + \theta\theta\overline{\theta\lambda}(x) + \overline{\theta\theta}\theta\xi(x) + \theta\theta\overline{\theta\theta}d(x) \end{aligned} \tag{4.82}$$

In the supermultiplet, there are four complex scalar components A, m, n, d, a complex vector V_μ, two Weyl spinors ψ, ξ that belong to $(1/2, 0)$, and another two Weyl spinors $\overline{\chi}, \overline{\lambda}$ in $(0, 1/2)$. Altogether, there are 16 real bosonic and 16 fermionic degrees of freedom. The transformation laws for each component under SUSY operator can be obtained by comparing powers of θ and $\overline{\theta}$ in the equation.

$$\delta F = \delta_\eta F + \delta_{\overline{\eta}} F = \delta A(x) + \theta\delta\psi + \overline{\theta}\delta\overline{\chi} + \theta\theta\delta m + \cdots \tag{4.83}$$

where the left-hand side (l.h.s.) is defined by

$$\delta_\eta F = \eta^\alpha i Q_\alpha F, \qquad \delta_{\overline{\eta}} F = \overline{\eta}_{\dot\alpha} i\overline{Q}^{\dot\alpha} F \tag{4.84}$$

Q_α, $\overline{Q}^{\dot{\alpha}}$ are given in Eqs. (4.74) and we obtain

$$\begin{aligned}
\delta A &= \eta^\alpha \psi_\alpha + \overline{\eta}_{\dot{\alpha}} \overline{\chi}^{\dot{\alpha}} \\
\delta \psi_\alpha &= 2\eta_\alpha m + (\sigma^\mu_{\alpha\dot{\beta}} \overline{\eta}^{\dot{\beta}})\left[i\partial_\mu A + V_\mu\right] \\
\delta \overline{\chi}^{\dot{\alpha}} &= 2\overline{\eta}^{\dot{\alpha}} n + (\eta^\alpha \sigma_\alpha^{\mu\dot{\alpha}})\left[i\partial_\mu A - V_\mu\right] \\
\delta m &= \overline{\eta}_{\dot{\alpha}} \overline{\lambda}^{\dot{\alpha}} - \tfrac{i}{2}(\partial_\mu \psi^\alpha \sigma^\mu_{\alpha\dot{\alpha}} \overline{\eta}^{\dot{\alpha}}) \\
\delta n &= \eta^\alpha \xi_\alpha + \tfrac{i}{2}(\eta^\alpha \sigma^\mu_{\alpha\dot{\alpha}} \partial_\mu \overline{\chi}^{\dot{\alpha}}) \\
\delta V_\mu &= (\eta^\alpha \sigma_{\mu\alpha\dot{\alpha}} \overline{\lambda}^{\dot{\alpha}}) + (\xi^\alpha \sigma_{\mu\alpha\dot{\alpha}} \overline{\eta}^{\dot{\alpha}}) + \tfrac{i}{2}(\eta^\alpha \partial_\mu \psi_\alpha) - \tfrac{i}{2}(\partial_\mu \overline{\chi}_\alpha \overline{\eta}^{\dot{\alpha}}) \\
\delta \overline{\lambda}^{\dot{\alpha}} &= 2\overline{\eta}^{\dot{\alpha}} d - i(\eta^\alpha \sigma_\alpha^{\mu\dot{\alpha}})\partial_\mu m + \tfrac{1}{4}(\overline{\eta\sigma}^{\mu\nu})^{\dot{\alpha}}(\partial_\mu V_\nu - \partial_\nu V_\mu) \\
\delta \xi_\alpha &= 2\eta_\alpha d + i(\sigma^\mu_{\alpha\dot{\alpha}} \overline{\eta}^{\dot{\alpha}})\partial_\mu n - \tfrac{1}{4}(\eta\sigma^{\mu\nu})_\alpha(\partial_\mu V_\nu - \partial_\nu V_\mu) \\
\delta d &= -\tfrac{i}{2}(\partial_\mu \xi^\alpha \sigma^\mu_{\alpha\dot{\alpha}} \overline{\eta}^{\dot{\alpha}}) + \tfrac{i}{2}(\eta^\alpha \sigma^\mu_{\alpha\dot{\alpha}} \partial_\mu \overline{\lambda}^{\dot{\alpha}})
\end{aligned} \tag{4.85}$$

where relations Eqs. (4.47) and (4.52) are used in deriving the symmetry transformations. Note that the variation of the highest component [i.e., $\delta d(x)$] of the superfield is a spacetime total derivative whose significance will be discussed later.

Problem 4.2 *Derive Eqs. (4.85)*

As the components are reducible under SUSY operation, which means that not all of them are required to close the algebra, we impose a condition to make them irreducible. One widely used set of conditions is given by

$$\overline{D}_{\dot{\alpha}} \Phi_L(x, \theta, \overline{\theta}) = 0 \tag{4.86a}$$

which defines a chiral (left-handed) superfield Φ_L. Similarly, a chiral right-handed superfield can be defined by

$$D_\alpha \Phi_R(x, \theta, \overline{\theta}) = 0 \tag{4.86b}$$

The reason for naming them left- or right-handed will be clarified later by Eq. (4.90). In what follows, we mainly discuss the left-handed chiral superfields.

Equation (4.86a) can be solved in terms of new variables defined by

$$y^\mu = x^\mu + i\theta\sigma^\mu\overline{\theta} \tag{4.87}$$

It is easy to show that

$$\overline{D}_{\dot{\alpha}} y^\mu = \overline{D}_{\dot{\alpha}} \theta_\alpha = 0 \tag{4.88}$$

Therefore, any function of y, θ is a solution to $\overline{D}_{\dot{\alpha}} \Phi = 0$. That is

$$\Phi_L(x, \theta, \overline{\theta}) = \hat{\Phi}(y, \theta) = \hat{\Phi}(x^\mu + i\theta\sigma^\mu\overline{\theta}, \theta) \tag{4.89}$$

where $\hat{\Phi}(y, \theta)$ is any function of y, θ. By Taylor expansion, the field $\hat{\Phi}$ can be expressed as

$$\hat{\Phi}(y, \theta) = A(y) + \sqrt{2}\theta^{\alpha}\psi_{\alpha}(y) + \theta\theta F(y) \tag{4.90}$$

Therefore, the superfield satisfying the condition Eq. (4.86a) contains only two scalar fields A, F and a left-handed Weyl spinor ψ. This justifies the claimed nomenclature. Notice that the chiral superfield has dimension $[E]$. It follows that the field $F(x)$ has dimension $[E^2]$, which is an indication that it is not a conventional scalar field. The expression for $\Phi_L(x, \theta, \overline{\theta})$ can be derived by going back to the original variable x.

$$\begin{aligned}\Phi_L(x, \theta, \overline{\theta}) &= \hat{\Phi}(x + i\theta\sigma^{\mu}\overline{\theta}, \theta)\\ &= A(x + i\theta\sigma^{\mu}\overline{\theta}) + \sqrt{2}\theta^{\alpha}\psi_{\alpha}(x + i\theta\sigma^{\mu}\overline{\theta}) + \theta\theta F(x + i\theta\sigma^{\mu}\overline{\theta})\\ &= A(x) + i(\theta\sigma^{\mu}\overline{\theta})\partial_{\mu}A(x) - \tfrac{1}{2}(\theta\sigma^{\mu}\overline{\theta})(\theta\sigma^{\nu}\overline{\theta})\partial_{\mu}\partial_{\nu}A(x)\\ &\quad + \sqrt{2}\theta^{\alpha}\psi_{\alpha}(x) + \sqrt{2}i\theta^{\alpha}(\theta\sigma^{\mu}\overline{\theta})\partial_{\mu}\psi_{\alpha}(x) + \theta\theta F(x)\end{aligned} \tag{4.91}$$

Using

$$\begin{aligned}(\theta\sigma^{\mu}\overline{\theta})(\theta\sigma^{\nu}\overline{\theta}) &\overset{(4.52)}{=\!=} \tfrac{1}{2}g^{\mu\nu}\theta\theta\overline{\theta\theta}\\ \theta^{\alpha}(\theta\sigma^{\mu}\overline{\theta}) = \theta^{\alpha}\theta^{\beta}\sigma^{\mu}_{\beta\dot{\beta}}\overline{\theta}^{\dot{\beta}} &\overset{(4.36)}{=\!=} -\tfrac{1}{2}\varepsilon^{\alpha\beta}\theta\theta\sigma^{\mu}_{\beta\dot{\alpha}}\overline{\theta}^{\dot{\alpha}}\end{aligned} \tag{4.92}$$

the chiral superfield can be expressed as

$$\begin{aligned}\Phi_L(x, \theta, \overline{\theta}) &= A(x) + \sqrt{2}\theta^{\alpha}\psi_{\alpha}(x) + \theta\theta F(x)\\ &\quad + i(\theta\sigma^{\mu}\overline{\theta})\partial_{\mu}A(x) + \frac{i}{\sqrt{2}}\theta\theta\overline{\theta}_{\dot{\alpha}}\sigma^{\mu\alpha\dot{\alpha}}\partial_{\mu}\psi_{\alpha}(x) - \tfrac{1}{4}\theta\theta\overline{\theta\theta}\partial_{\mu}\partial^{\mu}A\end{aligned} \tag{4.93}$$

Variations under the super-translation of the component fields A, ψ, and F in the chiral superfield Φ_L follow by applying the formula Eq. (4.85) to Eq. (4.91) and for the infinitesimal η, $\overline{\eta}$ are given by

$$\delta_{\eta}A = \sqrt{2}\eta^{\alpha}\psi_{\alpha} \tag{4.94a}$$

$$\delta_{\eta}\psi_{\alpha} = \sqrt{2}\left[\eta_{\alpha}F + i(\sigma^{\mu}\overline{\eta})_{\alpha}\partial_{\mu}A\right] \tag{4.94b}$$

$$\delta_{\eta}F = \sqrt{2}i\overline{\eta\sigma}^{\mu}\partial_{\mu}\psi \tag{4.94c}$$

One sees that no fields other than A, ψ, and F appear in the transformed fields. Every component in Eq. (4.93) transforms to each other, and hence the SUSY algebra is closed by themselves. That is, the chiral left-handed superfield constitutes an irreducible representation. It also means that the field F is an indispensable member to have a consistent SUSY algebra. Its physical meaning, however, is not clear at this stage. It will be clarified when the Lagrangian is constructed and the equation of motion for F is derived.

Noting

$$D_{\alpha}z^{\mu} = D_{\alpha}\overline{\theta}_{\dot{\alpha}} = 0 \tag{4.95}$$

where

$$z^{\mu} = x^{\mu} - i\theta\sigma^{\mu}\overline{\theta} \tag{4.96}$$

the right-handed chiral superfields can be obtained similarly:

$$\Phi_R(x,\theta,\overline{\theta}) = \hat{\Phi}(z,\overline{\theta}) = \hat{\Phi}(x^\mu - i(\theta\sigma^\mu\overline{\theta}),\overline{\theta}) \tag{4.97a}$$

$$\hat{\Phi}_R(z,\overline{\theta}) = A^*(z) + \sqrt{2}\overline{\theta}_{\dot{\alpha}}\overline{\psi}^{\dot{\alpha}}(z) + \overline{\theta\theta}F^*(z) \tag{4.97b}$$

From the above equation, one sees that the right-handed chiral superfield is conjugate to the left-handed field. That is, a right-handed superfield is a complex conjugate of a left-handed superfield and contains only the complex conjugate of a scalar field and a right-handed Weyl field.

4.3.1
Products of Chiral Superfields

Products of chiral superfields are again chiral superfields. This can be seen by explicitly making products of two left-handed scalar chiral fields.

$$\begin{aligned}\Phi_i &= A_i(y) + \sqrt{2}\theta\psi_i(y) + \theta\theta F_i(y)\\ \Phi_j &= A_j(y) + \sqrt{2}\theta\psi_j(y) + \theta\theta F_j(y)\end{aligned} \tag{4.98}$$

where i, j denote any quantum numbers to distinguish one superfield from the other. It is straightforward to show that

$$\begin{aligned}\Phi_i\Phi_j &= \left[A_i + \sqrt{2}\theta\psi_i + \theta\theta F_i\right]\left[A_j + \sqrt{2}\theta\psi_j + \theta\theta F_j\right]\\ &= A_iA_j + \sqrt{2}\theta\left[A_i\psi_j + \psi_iA_j\right] + \theta\theta\left[A_iF_j + F_iA_j - 2\psi_i\psi_j\right]\end{aligned} \tag{4.99}$$

As this is again a function of y and θ only, it satisfies $\overline{D}_\alpha\left[\Phi_i\Phi_j\right] = 0$, which proves the statement. By induction, product of three left-handed chiral superfields is again a left-handed chiral superfield. For the sake of later reference, we write down the triple scalar potential.

$$\begin{aligned}\Phi_i\Phi_j\Phi_k = {} & A_iA_jA_k + \sqrt{2}\theta\left[A_iA_j\psi_k + A_i\psi_jA_k + \psi_iA_jA_k\right]\\ & +\theta\theta\left[A_iA_jF_k + A_iF_jA_k + F_iA_jA_k - 2(\psi_i\psi_jA_k + \psi_j\psi_kA_i + \psi_k\psi_iA_j)\right]\end{aligned} \tag{4.100}$$

Similar results hold for the conjugate, that is, right-handed chiral fields.

Note that products of left- and right-handed chiral superfields such as $\Phi_i^\dagger\Phi_j$ are in general neither chiral nor antichiral. A special case of $i = j$ will be treated later in the construction of a supersymmetric Lagrangian.

4.4
Vector Superfields

The vector field is defined by the reality condition

$$V(x,\theta,\overline{\theta}) = V^\dagger(x,\theta,\overline{\theta}) \tag{4.101}$$

and it is an irreducible representation of the SUSY as will be shown soon. It can also be Taylor-expanded in a similar form as Eq. (4.82). The reality condition imposes constraints on the component fields and it can be expressed as

$$\begin{aligned} V(x,\theta,\bar{\theta}) = {} & \varphi(x) + \theta\chi(x) + \bar{\theta}\bar{\chi}(x) + \theta\theta m(x) + \overline{\theta\theta} m^*(x) + (\theta\sigma^{\mu}\bar{\theta})V_{\mu}(x) \\ & + \theta\theta\bar{\theta}\bar{\lambda}(x) + \overline{\theta\theta}\theta\lambda(x) + \theta\theta\overline{\theta\theta}D(x) \end{aligned} \tag{4.102}$$

where φ, V_{μ}, D are real fields. For the sake of later convenience, we modify the component fields and redefine them by

$$\begin{aligned} V(x^{\mu},\theta^{\alpha},\bar{\theta}_{\dot{\alpha}}) = {} & \varphi(x) + \theta\chi(x) + \bar{\theta}\bar{\chi}(x) + \theta\theta m(x) + \overline{\theta\theta} m^*(x) + (\theta\sigma_{\mu}\bar{\theta})V^{\mu}(x) \\ & + \theta\theta\bar{\theta}\left[\bar{\lambda}(x) - \tfrac{i}{2}\bar{\sigma}^{\mu}\partial_{\mu}\chi\right] + \overline{\theta\theta}\theta\left[\lambda(x) - \tfrac{i}{2}\sigma^{\mu}\partial_{\mu}\bar{\chi}\right] + \theta\theta\overline{\theta\theta}\left[\tfrac{1}{2}D(x) - \tfrac{1}{4}\partial_{\mu}\partial^{\mu}\varphi(x)\right] \end{aligned} \tag{4.103}$$

A special example of the vector field is a product of left-handed chiral field and its conjugate $(\Phi_L)^{\dagger}\Phi_L$ because it satisfies the reality condition

$$\{(\Phi_L)^{\dagger}\Phi_L\}^{\dagger} = (\Phi_L)^{\dagger}\Phi_L \tag{4.104}$$

Another important example is the sum of a left-handed field and its conjugate, as

$$\left(\Phi_L + (\Phi_L)^{\dagger}\right)^{\dagger} = \Phi_L^{\dagger} + \Phi_L = \Phi_L + (\Phi_L)^{\dagger} = \Phi_L + \Phi_R \tag{4.105}$$

Using Eq. (4.93), it can be expressed as

$$\begin{aligned} \Phi_L + \Phi_R = {} & A(x) + A(x)^* + \sqrt{2}\theta\psi(x) + \sqrt{2}\bar{\theta}\bar{\psi}(x) \\ & + \theta\theta F(x) + \overline{\theta\theta}F^*(x) + i(\theta\sigma^{\mu}\bar{\theta})\partial_{\mu}[A(x) - A^*(x)] \\ & + \frac{i}{\sqrt{2}}\theta\theta\bar{\theta}^{\dot{\alpha}}(\bar{\sigma}^{\mu}_{\dot{\alpha}\alpha})\partial_{\mu}\psi^{\alpha}(x) - \frac{i}{\sqrt{2}}\overline{\theta\theta}\theta^{\alpha}(\sigma^{\mu}_{\alpha\dot{\alpha}})\partial_{\mu}\bar{\psi}^{\dot{\alpha}}(x) \\ & - \tfrac{1}{4}\theta\theta\overline{\theta\theta}\partial_{\mu}\partial^{\mu}(A + A^*) \end{aligned} \tag{4.106}$$

One notices that this combination of scalar superfields has the gradient $i\partial_{\mu}[A(x) - A^*(x)]$ as the vector component, and by adding $\Phi_L + \Phi_R$ to the vector superfield $V(x)$, one can obtain the familiar gauge transformation of the vector. Therefore, we define the supersymmetric generalization of the gauge transformation by

$$\begin{aligned} V(x,\theta,\bar{\theta}) \to V'(x,\theta,\bar{\theta}) & = V(x,\theta,\bar{\theta}) + \Phi_L + \Phi_L^{\dagger} \\ & \equiv V(x,\theta,\bar{\theta}) + i[\Lambda(x,\theta,\bar{\theta}) - \Lambda^{\dagger}(x,\theta,\bar{\theta})] \end{aligned} \tag{4.107}$$

As the added term should be free to change, $\Phi_L + \Phi_L^{\dagger}$ has a lot more freedom compared to the ordinary gauge transformation. Let us see how each component changes. Substituting Eqs. (4.102) and (4.106) in Eq. (4.107), we obtain

$$\begin{aligned} \varphi'(x) & = \varphi(x) + A(x) + A^*(x) \\ \chi'_{\alpha}(x) & = \chi_{\alpha}(x) + \sqrt{2}\psi_{\alpha}(x) \\ m'(x) & = m(x) + F(x) \\ V'_{\mu}(x) & = V_{\mu}(x) + i\partial_{\mu}[A(x) - A^*(x)] \\ \lambda'_{\alpha}(x) & = \lambda_{\alpha}(x) - \tfrac{i}{2}\sigma^{\mu}_{\alpha\dot{\alpha}}\partial_{\mu}\left(\bar{\chi}^{\dot{\alpha}} + \sqrt{2}\bar{\psi}^{\dot{\alpha}}\right) \\ \tfrac{1}{2}D'(x) & = \tfrac{1}{2}D(x) - \tfrac{1}{4}\partial_{\mu}\partial^{\mu}[\varphi(x) + A(x) + A^*(x)] \end{aligned} \tag{4.108}$$

We may choose a particular gauge such that $\varphi' = \chi' = m' = 0$. Imposing the conditions

$$\begin{aligned} \varphi(x) + A(x) + A^*(x) &= 0 \\ \chi_\alpha(x) + \sqrt{2}\psi_\alpha(x) &= 0 \\ m(x) + F(x) &= 0 \end{aligned} \tag{4.109}$$

the gauge transformed vector takes the form

$$\begin{aligned} &\varphi'(x) = \chi'(x) = m'(x) = 0 \\ &V'_\mu(x) = V_\mu(x) - 2\partial_\mu[\mathrm{Im}A(x)] \\ &\lambda'_\alpha(x) = \lambda_\alpha(x) \\ &D'(x) = D(x) \end{aligned} \tag{4.110}$$

That is, by the generalized gauge transformation, $\lambda(x)$ and $D(x)$ remain invariant but others can be changed. By choosing a gauge where $\varphi = \chi = m = 0$, the physical content of the vector field becomes clear. In this gauge, it consists of a real gauge field, a fermion, and a real scalar. This particular form of the vector superfield has been obtained by fixing the generalized gauge, which is referred to as the *Wess–Zumino gauge*. In this gauge, the number of degrees of freedom is 3 for the vector field,[4] 1 for D, and 4 for λ. The number of degrees of freedom for the fermion and the boson is equal as it should be. The SUSY operation closes within the three fields, and hence they constitute an irreducible representation as claimed.

The fermion $\lambda(x)$ is a supersymmetric partner of the vector field and is referred to as the *gaugino*. Notice that the gauge field belongs to adjoint representation of symmetries; hence, the gaugino has no counterpart distinguished by the chirality. The Wess–Zumino gauge is adopted to eliminate redundant degrees of freedom so that physical contents of the vector superfield are transparent. It is similar to the unitary gauge in the Standard Model (SM).

It is easy to compute the power of the vector superfields, which we denote as V_{WZ}. With $\mathrm{Im}A = 0$, we have

$$\begin{aligned} &V_{WZ}(x) = (\theta\sigma^\mu\overline{\theta})V_\mu(x) + \theta\theta\overline{\theta}\,\overline{\lambda}(x) + \overline{\theta}\,\overline{\theta}\theta\lambda(x) + \theta\theta\overline{\theta}\,\overline{\theta}\tfrac{1}{2}D(x) \\ &(V_{WZ})^2 = \tfrac{1}{2}\theta\theta\overline{\theta}\,\overline{\theta}V_\mu(x)V^\mu(x) \\ &(V_{WZ})^3 = 0 \end{aligned} \tag{4.111}$$

Notice that, unlike the chiral superfield that has dimension $[E]$, the vector superfield is dimensionless. It follows that the field D has dimension $[E^2]$ just like F. The role of the field $D(x)$ is similar to F in the chiral field and its physical meaning will again be clarified using the equation of motion later.

---

[4] This is because we still have the freedom to change the Im A to reduce one extra degree of freedom.

4.4.1
Field Strength

Having given the vector field and its gauge transformation, we can define the supersymmetric generalization of the field strength, a gauge covariant tensor $F_{\mu\nu}$ as a component of the following quantities:

$$W_\alpha = \frac{1}{4}\overline{DD}D_\alpha V, \qquad \overline{W}_{\dot{\alpha}} = \frac{1}{4}DD\overline{D}_{\dot{\alpha}}V \tag{4.112}$$

where $\overline{DD} = \overline{D}_{\dot{\alpha}}\overline{D}^{\dot{\alpha}}$, $DD = D^\alpha D_\alpha$. We can show that W_α and $\overline{W}_{\dot{\alpha}}$ are left- and right-handed chiral superfields. This is because W_α satisfies

$$\overline{D}_{\dot{\alpha}}W_\alpha = \frac{1}{4}\overline{D}_{\dot{\alpha}}\overline{D}^2 D_\alpha V = 0 \tag{4.113}$$

where we used $(\overline{D}^3)_{\dot{\alpha}} = 0$ [Eq. (4.77)]. Equation (4.113) is nothing but a defining equation for the chiral field. The gauge invariance of W_α can be shown if $W'_\alpha = W_\alpha$. By applying the gauge transformation Eq. (4.107),

$$\begin{aligned} W'_\alpha &= \frac{1}{4}\overline{DD}D_\alpha V' = \frac{1}{4}\overline{DD}D_\alpha(V + \Phi_L + \Phi_R) \\ &= \frac{1}{4}\overline{DD}D_\alpha V + \frac{1}{4}\overline{D}_{\dot{\beta}}\{\overline{D}^{\dot{\beta}} D_\alpha\}\Phi_L - \frac{1}{4}\overline{D}_{\dot{\beta}}D_\alpha\overline{D}^{\dot{\beta}}\Phi_L + \frac{1}{4}\overline{DD}D_\alpha\Phi_R \end{aligned} \tag{4.114a}$$

The last two terms vanish because of Eqs. (4.86). The second term is modified to

$$W'_\alpha \overset{(4.77\text{d})}{=\!=} W_\alpha + \frac{1}{4}\overline{D}_{\dot{\beta}}(-2i\sigma_\alpha^{\mu\,\dot{\beta}}\partial_\mu)\Phi_L = W_\alpha - \frac{i}{2}(\sigma_\alpha^{\mu\,\dot{\beta}}\partial_\mu)\overline{D}_{\dot{\beta}}\Phi_L = W_\alpha \tag{4.114b}$$

where the penultimate equality follows from commutativity of $\overline{D}_{\dot{\beta}}$ and P_μ. Proof for $\overline{W}_{\dot{\alpha}}$ can be done similarly. Using explicit representation for D_α, $\overline{D}_{\dot{\alpha}}$ [Eqs. (4.76)], we can calculate W_α in the Wess–Zumino gauge. The calculation is made a bit easier by working with variables $y^\mu = x^\mu + i(\theta\sigma^\mu\overline{\theta})$ or $z^\mu = x^\mu - i(\theta\sigma^\mu\overline{\theta})$. In the y- (or z-) coordinate system, which we call L- (or R-) representation, the covariant derivatives assume simpler forms

$$D_\alpha^{(L)} \equiv \nabla_\alpha + 2i\sigma^\mu_{\alpha\dot{\alpha}}\overline{\theta}^{\dot{\alpha}}(\partial_y)_\mu, \qquad \overline{D}_{\dot{\alpha}}^{(L)} \equiv -\overline{\nabla}_{\dot{\alpha}} \tag{4.115a}$$

$$D_\alpha^{(R)} \equiv \nabla_\alpha, \qquad \overline{D}_{\dot{\alpha}}^{(R)} \equiv -\overline{\nabla}_{\dot{\alpha}} - 2i\theta^\alpha\sigma^\mu_{\alpha\dot{\alpha}}(\partial_z)_\mu \tag{4.115b}$$

where the spacetime derivatives are $(\partial_y)_\mu = \partial/\partial y^\mu$ and $(\partial_z)_\mu = \partial/\partial z^\mu$. Then, we have

$$\begin{aligned} W_\alpha(y,\theta,\overline{\theta}) &= \tfrac{1}{4}\overline{D}^{(L)}\overline{D}^{(L)}D_\alpha^{(L)}V_{WZ}(y - i(\theta\sigma^\mu\overline{\theta}),\theta,\overline{\theta}) \\ &= \tfrac{1}{4}\left(\overline{\nabla}_{\dot{\alpha}}\overline{\nabla}^{\dot{\alpha}}\right)\left(\nabla_\alpha + 2i\sigma^\mu_{\alpha\dot{\beta}}\overline{\theta}^{\dot{\beta}}(\partial_y)_\mu\right)V_{WZ}(y^\mu - i\theta\sigma^\mu\overline{\theta},\theta,\overline{\theta}) \\ \overline{W}_{\dot{\alpha}}(z,\theta,\overline{\theta}) &= \tfrac{1}{4}D^{(R)}D^{(R)}\overline{D}_{\dot{\alpha}}^{(R)}V_{WZ}(z + i(\theta\sigma^\mu\overline{\theta}),\theta,\overline{\theta}) \\ &= \tfrac{1}{4}\left(\nabla^\alpha\nabla_\alpha\right)\left(-\overline{\nabla}_{\dot{\alpha}} - 2i\theta^\alpha\sigma^\mu_{\alpha\dot{\alpha}}(\partial_z)_\mu\right)V_{WZ}(z^\mu + i\theta\sigma^\mu\overline{\theta},\theta,\overline{\theta}) \end{aligned} \tag{4.116}$$

To carry out the algebra, one first expands $V_{ZW}(y - i\theta\sigma\overline{\theta}, \theta, \overline{\theta})$. Referring to Eq. (4.111)

$$V_{WZ}\left(y - i(\theta\sigma\overline{\theta}), \theta, \overline{\theta}\right) = (\theta\sigma^{\mu}\overline{\theta})V_{\mu} + \theta\theta\overline{\theta\lambda} + \overline{\theta\theta}\theta\lambda + \frac{1}{2}\theta\theta\overline{\theta\theta}(D - i\partial_{\mu}V^{\mu}) \quad (4.117)$$

and apply $D_{\alpha}^{(L)}$, and so on. A straightforward calculation gives

$$\begin{aligned}
W_{\alpha}(y, \theta, \overline{\theta}) &= \lambda_{\alpha}(y) + \theta_{\alpha}D(y) + (\sigma^{\mu\nu}_{\alpha\beta}\theta^{\beta})F_{\mu\nu}(y) - i\theta\theta\sigma^{\mu}_{\alpha\dot{\beta}}\partial_{\mu}\overline{\lambda}^{\dot{\beta}}(y) \\
\overline{W}_{\dot{\alpha}}(z, \theta, \overline{\theta}) &= \overline{\lambda}_{\dot{\alpha}}(z) + \overline{\theta}_{\dot{\alpha}}D(z) + (\overline{\theta}_{\dot{\beta}}\overline{\sigma}^{\mu\nu\dot{\beta}}{}_{\dot{\alpha}})F_{\mu\nu}(z) + i\overline{\theta\theta}\partial_{\mu}\lambda^{\beta}(z)\sigma^{\mu}_{\beta\dot{\alpha}} \\
F_{\mu\nu} &= \partial_{\mu}V_{\nu} - \partial_{\nu}V_{\mu}, \quad \sigma^{\mu\nu} = \tfrac{i}{2}\left(\sigma^{\mu}\overline{\sigma}^{\nu} - \sigma^{\nu}\overline{\sigma}^{\mu}\right), \quad \overline{\sigma}^{\mu\nu} = \tfrac{i}{2}\left(\overline{\sigma}^{\mu}\sigma^{\nu} - \overline{\sigma}^{\nu}\sigma^{\mu}\right)
\end{aligned} \quad (4.118)$$

Equations (4.118) contain the field strength as desired.

4.5
Action

4.5.1
SUSY Invariant Action

The action in superspace is given by

$$\begin{aligned}
S &= \int d^4x \mathcal{L}(x) \\
\mathcal{L}(x) &= \int d^2\theta d^2\overline{\theta}\, \mathscr{L}\left[\Phi(x, \theta, \overline{\theta}), V(x, \theta, \overline{\theta})\right]
\end{aligned} \quad (4.119)$$

where the integration includes over four-dimensional super-coordinates $(\theta^1, \theta^2, \overline{\theta}_{\dot{1}}, \overline{\theta}_{\dot{2}})$ as well as over spacetime. For a while, we omit the vector field for simplicity. We recall that integration over a Grassmann variable η is given by

$$\int d\eta = 0, \qquad \int d\eta\eta = 1 \quad (4.120)$$

Recalling $\theta\theta = \theta^{\alpha}\theta_{\alpha} = \theta^{\alpha}\varepsilon_{\alpha\beta}\theta^{\beta} = -2\theta^1\theta^2$, we have

$$\begin{aligned}
\int d^2\theta\theta\theta &\equiv c \int d\theta^1 d\theta^2(-2\theta^1\theta^2) = 2c \int d\theta^1 d\theta^2\theta^2\theta^1 = 2c \int d\theta^1\theta^1 \\
&= 2c
\end{aligned} \quad (4.121)$$

By requiring the measure to satisfy the same relation as Eq. (4.120), we have $c = 1/2$. Therefore, the measure $d^2\theta$ can be defined as

$$d^2\theta = -\frac{1}{4}\varepsilon_{\alpha\beta}d\theta^{\alpha}d\theta^{\beta}, \qquad \int d^2\theta\, \theta\theta = 1 \quad (4.122)$$

Defining the measure $d^2\overline{\theta}$ similarly, we have

$$d^2\overline{\theta} = -\frac{1}{4}\varepsilon^{\dot{\alpha}\dot{\beta}}d\overline{\theta}_{\dot{\alpha}}d\overline{\theta}_{\dot{\beta}}, \qquad \int d^2\overline{\theta}\, \overline{\theta\theta} = 1 \quad (4.123)$$

Let us consider the integration of a superfield that is at most quadratic in θ.

$$f(\theta) = A + \theta^{\alpha}B_{\alpha} + C\theta\theta \quad (4.124)$$

From Eq. (4.120), we have

$$\int d^2\theta f(\theta) = \int d^2\theta\,(A + \theta^\alpha B_\alpha + C\theta\theta) = C \tag{4.125}$$

with a similar equation for $\int d^2\overline{\theta}\, g(\overline{\theta})$. Consequently, the only nonvanishing terms in Eq. (4.119) are those proportional to $\theta\theta\overline{\theta\theta}$. When the integrand contains a function of θ or $\overline{\theta}$ only, Eq. (4.119) is to be understood as

$$\mathcal{L}(x) = \int d^2\theta d^2\overline{\theta}\Big[\mathscr{L}_L[\Phi_L(x,\theta)]\delta(\overline{\theta\theta}) + \mathscr{L}_R[\Phi_R(x,\overline{\theta})]\delta(\theta\theta) + \mathscr{L}\left[\Phi(x,\theta,\overline{\theta})\right]\Big] \tag{4.126}$$

where $\delta(\theta\theta) = \theta\theta$ is a δ function whose integral gives unity. That is, the superspace integral picks up the highest order component of the Lagrangian made of superfields. All other terms vanish upon integration.

The action that is invariant under the SUSY should satisfy the relation

$$\delta_S S = \int d^4x\,\delta_S\mathcal{L}(x) = 0 \tag{4.127}$$

This means that the SUSY action should give $\delta_S\mathcal{L} = 0$ up to a total divergence. We have seen that the highest component of a superfield always gives a total divergence upon the symmetry operation [Eqs. (4.85) and (4.94)]. If they also have the proper Lorentz structure, like those of kinetic energy, a mass term or a potential, they are candidates of the Lagrangian. By demonstrating that the superspace integration of the action picks up only the highest component of the superfields, we have just proved that the action defined by Eq. (4.119) is SUSY-invariant.

In the following, we consider building a Lagrangian using irreducible representations of the superfields, namely the chiral fields and the vector fields. The vector fields are real. The chiral field contains either θ only (left-handed) or $\overline{\theta}$ only (right-handed). As the Lagrangian has to be Hermitian and the right-handed chiral fields are obtained by taking the Hermitian conjugate of the left-handed fields, we can separate the Lagrangian into two:

$$\begin{aligned}\mathcal{L}(x) &= \int\left[\left\{\int d^2\theta\mathscr{L}_L + h.c.\right\} + \int d^2\theta d^2\overline{\theta}\mathscr{L}\right]\\ &= \int d^2\theta d^2\overline{\theta}\left[\left\{\mathscr{L}_L\delta(\theta\theta) + h.c.\right\} + \mathscr{L}\right]\end{aligned} \tag{4.128}$$

where $\delta(\theta\theta)$ is the δ function. From now on, we omit to attach the suffix "L" to the chiral superfield and simply denote it as Φ unless otherwise specified.

We have already seen that products of chiral fields contain mass terms [Eq. (4.99)] and Yukawa interactions [Eq. (4.100)], which are candidates to be included in the chiral part of the Lagrangian Eq. (4.128). Our next task is to look for a superfield that contains the kinetic energy.

4.5.2
Kinetic Energy of Chiral Superfield

A second order or square of the derivatives is necessary to construct the kinetic energy part of the Lagrangian. We have seen already that chiral fields or their products contain no derivatives at least in the L or R representation.

Let us, instead, consider the product of a left-chiral field and its conjugate. As the left-handed field is in the L representation and its conjugate is in the R representation, we need to unify the representation to handle their product. Expressing the left-chiral field explicitly in the x-representation, we have

$$\begin{aligned}\Phi(x,\theta,\overline{\theta}) \overset{(4.93)}{=\!=} & \underbrace{A}_{(1)} + \underbrace{\sqrt{2}\theta^{\alpha}\psi_{\alpha}(x)}_{(2)} + \underbrace{\theta\theta F(x)}_{(3)} + \underbrace{i(\theta^{\alpha}\sigma^{\mu}_{\alpha\dot\alpha}\overline{\theta}^{\dot\alpha})\partial_{\mu}A(x)}_{(4)} \\ & + \underbrace{\frac{i}{\sqrt{2}}\theta\theta\overline{\theta}^{\dot\alpha}\left(\sigma^{\mu}_{\alpha\dot\alpha}\partial_{\mu}\psi^{\alpha}(x)\right)}_{(5)} - \underbrace{\frac{1}{4}\theta\theta\overline{\theta\theta}\partial_{\mu}\partial^{\mu}A}_{(6)} \end{aligned} \tag{4.129}$$

Taking the Hermitian conjugate and noting the Hermitian nature of the Grassmann variables, that is, $(\theta^{\alpha}\overline{\theta}^{\dot\beta})^{*} = (\overline{\theta}^{\dot\beta})^{*}(\theta^{\alpha})^{*} = \theta^{\beta}\overline{\theta}^{\dot\alpha}$, we have

$$\begin{aligned}\Phi^{\dagger}(x,\theta,\overline{\theta}) = & \underbrace{A^{*}}_{(7)} + \underbrace{\sqrt{2}\overline{\theta}_{\dot\alpha}\overline{\psi}^{\dot\alpha}(x)}_{(8)} + \underbrace{\overline{\theta\theta}F^{*}(x)}_{(9)} \\ & - \underbrace{i(\theta^{\alpha}\sigma^{\mu}_{\alpha\dot\alpha}\overline{\theta}^{\dot\alpha})\partial_{\mu}A^{*}(x)}_{(10)} + \underbrace{\frac{i}{\sqrt{2}}\overline{\theta\theta}\theta^{\alpha}\left(\sigma^{\mu}_{\alpha\dot\alpha}\partial_{\mu}\overline{\psi}^{\dot\alpha}(x)\right)}_{(11)} - \underbrace{\frac{1}{4}\theta\theta\overline{\theta\theta}\partial_{\mu}\partial^{\mu}A^{*}}_{(12)} \end{aligned} \tag{4.130}$$

We only need to consider the coefficients of $\theta\theta\overline{\theta\theta}$ in $\Phi^{\dagger}\Phi$. The fermionic kinetic energy terms arise from

$$\begin{aligned}&(2)\times(11)+(5)\times(8) \\ &= \sqrt{2}\theta^{\beta}\psi_{\beta}(x)\frac{i}{\sqrt{2}}\overline{\theta\theta}\theta^{\alpha}\left(\sigma^{\mu}_{\alpha\dot\alpha}\partial_{\mu}\overline{\psi}^{\dot\alpha}(x)\right) + \frac{i}{\sqrt{2}}\theta\theta\overline{\theta}^{\dot\alpha}\left(\sigma^{\mu}_{\alpha\dot\alpha}\partial_{\mu}\psi^{\alpha}(x)\right)\sqrt{2}\overline{\psi}_{\dot\beta}(x)\overline{\theta}^{\dot\beta} \\ &= \frac{i}{2}\theta\theta\overline{\theta\theta}\left[-\psi^{\alpha}\sigma^{\mu}_{\alpha\dot\alpha}\partial_{\mu}\overline{\psi}^{\dot\alpha}+\partial_{\mu}\psi^{\alpha}\sigma^{\mu}_{\alpha\dot\alpha}\overline{\psi}^{\dot\alpha}\right] = \frac{i}{2}\theta\theta\overline{\theta\theta}\left[\partial_{\mu}\overline{\psi}^{\dot\alpha}\overline{\sigma}^{\mu}_{\dot\alpha\alpha}\psi^{\alpha} - \overline{\psi}^{\dot\alpha}\overline{\sigma}^{\mu}_{\dot\alpha\alpha}\partial_{\mu}\psi^{\alpha}\right] \\ &= -\frac{i}{2}\left[\overline{\psi}^{\dot\alpha}\overline{\sigma}_{\mu\,\dot\alpha\alpha}\overset{\leftrightarrow}{\partial_{\mu}}\psi^{\alpha}\right]\theta\theta\overline{\theta\theta} = -\frac{i}{2}(\overline{\psi}\overline{\sigma}^{\mu}\overset{\leftrightarrow}{\partial_{\mu}}\psi)\,\theta\theta\overline{\theta\theta}\end{aligned} \tag{4.131}$$

The bosonic kinetic energy terms are

$$
\begin{aligned}
&(1)\times(12)+(6)\times(7)+(3)\times(9)+(4)\times(10)\\
&=-\frac{1}{4}\left(A\partial_\mu\partial^\mu A^*+A^*\partial_\mu\partial^\mu A\right)\theta\theta\overline{\theta\theta}+FF^*\theta\theta\overline{\theta\theta}+(\theta\sigma^\mu\overline{\theta})(\theta\sigma^\nu\overline{\theta})\partial_\mu A\partial_\nu A^*\\
&=\left[\left\{\frac{1}{2}\partial_\mu A^*\partial^\mu A-\frac{1}{4}\partial_\mu(A^*\partial^\mu A+\partial^\mu A^*A)\right\}+FF^*+\frac{1}{2}g^{\mu\nu}\partial_\mu A^*\partial_\nu A\right]\theta\theta\overline{\theta\theta}\\
&=\left[\partial_\mu A^*\partial^\mu A+FF^*+\text{divergence term}\right]\theta\theta\overline{\theta\theta}
\end{aligned}
\tag{4.132}
$$

where Eq. (4.52) was used in deriving the last term of the penultimate line. Combining Eqs. (4.131) and (4.132), the Lagrangian for the kinetic energy part of the chiral fields becomes

$$
\mathcal{L}_{\text{chiral, KE}}(x)=\int d^2\theta d^2\overline{\theta}\Phi^\dagger\Phi=\partial_\mu A^*\partial^\mu A-\tfrac{1}{2}\left(\overline{\psi}\overline{\sigma}^\mu\, i\overleftrightarrow{\partial_\mu}\,\psi\right)+FF^* \tag{4.133}
$$

The first two terms represent the kinetic energy of a scalar field and its super-partner left-handed Weyl field, but the Lagrangian also contains a term expressed as FF^*. The meaning of the last term will be clarified soon.

4.5.3
Superpotential

Before discussing the physical meaning of the field F, we need to organize the Lagrangian to include the mass and potential terms. If one looks at the highest term of $\Phi_i\Phi_j$ in Eq. (4.99), it contains the fermion mass term $\sim\psi_i\psi_j$. So it is a candidate for the mass term. Similarly, the highest term of $\Phi_i\Phi_j\Phi_k$ [Eq. (4.100)] contains Yukawa interaction $\sim\psi_i\psi_j A_i$, which was necessary in the SM. As the integration of $d^2\theta d^2\overline{\theta}$ in the expression for the action automatically picks up the highest term in the component expansion of the superfields, the Lagrangian for the chiral scalar superfields will have a generic form

$$
\mathcal{L}_{\text{chiral}}(x)=\mathcal{L}_{\text{KE}}(x)+\mathcal{L}_{\text{SP}}(x)=\int d^2\theta d^2\overline{\theta}\left[\mathscr{L}_{\text{KE}}+\left(\delta(\overline{\theta}^2)W[\Phi]+h.c.\right)\right] \tag{4.134}
$$

$$
\mathcal{L}_{\text{KE}}=\sum_i\left[\partial_\mu A_i^*\partial^\mu A_i-\frac{1}{2}\left(\overline{\psi}_i{}^{\dot{\alpha}}\overline{\sigma}^\mu_{\dot{\alpha}\alpha}\, i\overleftrightarrow{\partial_\mu}\,\psi_i{}^\alpha\right)+F_iF_i^*\right] \tag{4.135}
$$

$$
W[\Phi]=\sum_{ijk}\left[t_i\Phi_i+\frac{1}{2}m_{ij}\Phi_i\Phi_j+\frac{1}{3}y_{ijk}\Phi_i\Phi_j\Phi_k\right] \tag{4.136}
$$

Contents of the third line are referred to as the *superpotential*. It is a polynomial of up to the third powers of the superfield. The renormalizability condition prohibits higher powers of Φ than the third, but a linear term is permissible.[5] Notice that m_{ij}'s has dimension E, whereas y_{ijk}s are dimensionless. They are symmetric under exchange of i, j, k's. Since the superpotential does not allow superfields of opposite chirality which are complex conjugate of the original superfields, it is an analytic function of the chiral superfields regarded as complex numbers.[6] The superpotential in its form given by Eq. (4.136) is hard to grasp its physical contents. There is, however, a simple formula to convert it to a function of conventional fields.

---

[5] As will become clear later, the cubic power of the superpotential contains quartic power of the scalar potential after the field F is re-expressed in terms of physical fields using the equation of motion. Terms in the Lagrangian with dimension >4 induces unrenormalizable interactions.

Theorem 1

$$\int d^2\theta W[\Phi] = \frac{\partial W(A)}{\partial A(x)} F - \frac{1}{2}\frac{\partial^2 W(A)}{\partial A^2}\psi\psi \tag{4.137}$$

where $W(A)$ is obtained from $W[\Phi]$ by replacing Φ with A.

Proof: We will prove it for the case where the superpotential is given by Eq. (4.136). For this case, one obtains

$$\frac{\partial W(A)}{\partial A_i(x)} F_i - \frac{1}{2}\frac{\partial^2 W(A)}{\partial A_i \partial A_j}\psi_i\psi_j = t_i F_i + m_{ij} F_i A_j + y_{ijk} A_i A_j F_k - \frac{1}{2} m_{ij}\psi_i\psi_j - y_{ijk}\psi_i\psi_j A_k \tag{4.138}$$

We will show that honest calculation integrating the superpotential over the superspace coordinates gives exactly the same expression as Eq. (4.138). We recall that all the left-handed chiral superfields have a simple form given by Eq. (4.90) in the L representation, that is, if they are expressed in the variable y. Fortunately, the integral over the superspace coordinates picks up only the highest component of the integrand. Then, we are free to change the variable y or z to x, because, for example,

$$\begin{aligned}\theta\theta D(y) &= \theta\theta D(x + i\theta\sigma\overline{\theta}) = \theta\theta\left[D(x) + i(\theta\sigma^{\mu}\overline{\theta})\partial_{\mu} D(x) - (\theta\sigma^{\mu}\overline{\theta})(\theta\sigma^{\nu}\overline{\theta})\partial_{\mu}\partial_{\nu} D(x)\right] \\ &= \theta\theta D(x)\end{aligned} \tag{4.139}$$

Therefore, substituting Eqs. (4.99) and (4.100) in Eq. (4.136), integration over superspace coordinates removes all terms except the highest order components. Then, changing the variables to x, we obtain

$$\mathcal{L}_{\text{SP}}(x) = \left[\left[t_k + m_{ik}A_i + y_{ijk}A_iA_j\right]F_k - \tfrac{1}{2}m_{ij}\psi_i\psi_j - y_{ijk}\psi_i\psi_j A_k\right] \tag{4.140}$$

The equation has reproduced Eq. (4.138). ■

---

[6] Mathematically, it is referred to as a holomorphic function of complex variables which is differentiable infinite times.

4.5.4
Lagrangian of the Chiral Fields

The Lagrangian for the chiral field is given by sum of kinetic energy term Eq. (4.133) and the superpotential.

$$\mathcal{L}_{\text{chiral}}(x) = \mathcal{L}_{\text{KE}} + \mathcal{L}_{\text{SP}} \tag{4.141a}$$

$$\mathcal{L}_{\text{KE}}(x) = \sum_i \left[\partial_\mu A_i^*(x)\partial^\mu A(x)_i - \frac{1}{2}\left\{\overline{\psi}_i(x)\overline{\sigma}^\mu \, i\overleftrightarrow{\partial_\mu} \, \psi_i(x)\right\} + |F_i|^2\right] \tag{4.141b}$$

$$\mathcal{L}_{\text{SP}}(x) = \sum_{ijk} \left[W_i F_i - \frac{1}{2}\sum_{j,k} W_{jk}\psi_j\psi_k + h.c.\right] \tag{4.141c}$$

$$W_i = \frac{\partial W(A)}{\partial A_i}, \qquad W_{ij} = \frac{\partial^2 W(A)}{\partial A_i \partial A_j} \tag{4.141d}$$

$$W(A) = \sum_{ijk} \left[t_i A_i + \frac{1}{2}m_{ij}A_iA_j + \frac{1}{3}\gamma_{ijk}A_iA_jA_k\right] \tag{4.141e}$$

Note that $W(A)$ is no longer a function of superfields but a function of the conventional scalar field $A(x)$. Given a model for superpotential $W(A)$, Eq. (4.137) gives a recipe to calculate the potential that works among the scalar fields.

For the superpotential given by Eq. (4.141e), the explicit form of the Lagrangian is given by

$$\begin{aligned}\mathcal{L}_{\text{chiral}}(x) = &\sum_i \left[\partial_\mu A_i^*(x)\partial^\mu A(x)_i - \frac{1}{2}\left\{\overline{\psi}_i(x)\overline{\sigma}^\mu \, i\overleftrightarrow{\partial_\mu} \, \psi_i(x)\right\} + |F_i|^2\right] \\ &+ \sum_{ijk}\left[\left(t_k + m_{ik}A_i + \gamma_{ijk}A_iA_j\right)F_k - \frac{1}{2}m_{ij}\psi_i\psi_j - \gamma_{ijk}\psi_i\psi_jA_k + h.c.\right]\end{aligned} \tag{4.142}$$

This is the so-called offline expression of the SUSY invariant Lagrangian.

Auxiliary field The field F_i has dimension E^2, as can be seen from Eq. (4.142). It does not have the kinetic energy part in the Lagrangian. This means that F_i does not propagate in spacetime and hence is not a dynamical object. Using the equation of motion, they can be rewritten in terms of the other dynamical objects. Therefore, the fields F_i are auxiliary fields and are not physical. Using the Euler equation for F_is

$$\partial_\mu \frac{\delta\mathcal{L}}{\delta(\partial_\mu F_i)} - \frac{\partial\mathcal{L}}{\partial F_i} = 0 \tag{4.143}$$

For free fields, that is, with no superpotential, the equation of motion gives the trivial result $F = 0$. If one includes interactions, the equation of motion results in

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial F_i(x)} = 0 \quad &\rightarrow \quad F_i^*(x) = -\left[t_i + m_{ij}A_j(x) + \gamma_{ijk}A_j(x)A_k(x)\right] \\ \frac{\partial\mathcal{L}}{\partial F_i^*(x)} = 0 \quad &\rightarrow \quad F_i(x) = -\left[t_i^* + m_{ij}^*A_j^*(x) + \gamma_{ijk}^*A_j^*(x)A_k^*(x)\right]\end{aligned} \tag{4.144}$$

Substituting Eqs. (4.144) in (4.142), we obtain

$$
\begin{aligned}
S_{\text{chiral}} = \int d^4x \Bigg[& \left\{ \partial_\mu A_i^*(x)\partial^\mu A(x)_i - \tfrac{1}{2}\left\{\overline{\psi}_i(x)\overline{\sigma}^\mu\, i\overleftrightarrow{\partial_\mu}\, \psi_i(x)\right\} \right\} \\
& - \left\{ \tfrac{1}{2} m_{ij}\psi_i(x)\psi_j(x) + y_{ijk}\psi_i(x)\psi_j(x)A_k(x)\ + (h.c.) \right\} - V_F[A_i(x),\, A_i^*(x)] \\
V_F[A_i(x),\, A_i^*(x)] = {} & \textstyle\sum_i F_i(x)F_i^*(x) = \sum_{ijk} \left| t_i + m_{ij}A_j(x) + y_{ijk}A_j(x)A_k(x) \right|^2
\end{aligned}
\tag{4.145}
$$

Now the physical content of the superpotential is clear. This is the Lagrangian for the scalar field $A_i(x)$ and the fermion field $\psi_i(x)$, having mass m (after diagonalizing the mass matrix $[m]_{ij}$), the Yukawa interaction between them and interactions among the scalars $V(A_i,\, A_i^*)$, which is commonly referred to as the *F-term*. Now everything looks familiar and physical. But notice that we realized the physically transparent expression at the expense of going on-shell. If one wants the SUSY operation off-shell as well as on-shell, one has to retain the auxiliary field. As the path integral formalism requires off-shell as well as on-shell fields [see Chapter 11 of [1]], proper treatments of the quantized fields cannot be realized without off-shell contribution of the fields.

To complete the action, we need to include the vector fields and their interaction.

4.5.5
Kinetic Energy of Vector Field

We first construct the kinetic energy part of the Lagrangian for the vector field. We have already seen that the field strength $F_{\mu\nu} = \partial_\mu V_\nu - \partial_\nu V_\mu$ resides in W_α and in $\overline{W}_{\dot\alpha}$. As $W^\alpha W_\alpha$ is a Lorentz scalar, we may try to make $W^\alpha W_\alpha$ and $\overline{W}_{\dot\alpha}\overline{W}^{\dot\alpha}$. After straightforward calculation, we have

$$
\begin{aligned}
W^\alpha W_\alpha = {} & \lambda^2(y) + 2D(y)\lambda(y)\theta + 2\lambda(y)\sigma^{\mu\nu}\theta F_{\mu\nu}(y) \\
& + \theta\theta\left[D^2(y) - 2i\lambda(y)\sigma^\mu\partial_\mu\overline{\lambda}(y) - \frac{1}{2}F_{\mu\nu}(y)F^{\mu\nu}(y) - \frac{i}{2}F_{\mu\nu}(y)\tilde{F}^{\mu\nu}(y) \right]
\end{aligned}
\tag{4.146a}
$$

$$
\begin{aligned}
\overline{W}_{\dot\alpha}\overline{W}^{\dot\alpha} = {} & \overline{\lambda}^2(z) + 2D(z)\overline{\lambda}(z)\overline{\theta} - 2\overline{\lambda}(z)\overline{\sigma}^{\mu\nu}\overline{\theta}F_{\mu\nu}(z) \\
& + \overline{\theta\theta}\left[D^2(z) - 2i\partial_\mu\lambda(z)\sigma^\mu\overline{\lambda}(z) - \frac{1}{2}F_{\mu\nu}(z)F^{\mu\nu}(z) + \frac{i}{2}F_{\mu\nu}(z)\tilde{F}^{\mu\nu}(z) \right]
\end{aligned}
\tag{4.146b}
$$

$$
\tilde{F}^{\mu\nu} = \frac{1}{2}\varepsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}, \qquad y^\mu = x^\mu + i(\theta\sigma^\mu\overline{\theta}), \qquad z^\mu = x^\mu - i(\theta\sigma^\mu\overline{\theta})
\tag{4.146c}
$$

Although they are given in L- and R-representations, respectively, we can change their variables y and z to x as stated before because the action picks up only the highest components. Therefore,

$$
\begin{aligned}
\mathcal{L}_{\text{vector, KE}} &= \int d^2\theta d^2\overline{\theta}\, \tfrac{1}{4}\left(W^\alpha W_\alpha + \overline{W}_{\dot\alpha}\overline{W}^{\dot\alpha} \right) \\
&= \left[\tfrac{1}{2}D^2 - \tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} - i\lambda\sigma^\mu\partial_\mu\overline{\lambda} \right]
\end{aligned}
\tag{4.147}
$$

where the factor 1/4 is used to obtain proper normalization. One notices that this is the Lagrangian of a massless gauge boson, a massless fermion that is a super partner of the gauge boson, and an auxiliary field D. Thus, we have identified the kinetic energy part of the Lagrangian for the gauge field. The additional field D can be treated in a manner similar to the F term in the chiral fields and can be rewritten in terms of the scalar fields after they are built into the total Lagrangian including the chiral fields.

4.6
Gauge Interaction

4.6.1
Global *U*(1) Transformation

To connect the gauge field with matter fields, let us first consider a global $U(1)$ gauge interaction. We define the global gauge transformation of a scalar superfield by

$$\Phi \rightarrow \Phi'_i = e^{-iq_i\alpha}\Phi_i \tag{4.148}$$

where q_i is the $U(1)$ charge of the superfield Φ and α is a real constant. Under the transformation, the Lagrangian changes to

$$\begin{aligned}\mathscr{L}' &= \left(t_i\Phi'_i + \frac{1}{2}m_{ij}\Phi'_i\Phi'_j + \frac{1}{3}\gamma_{ijk}\Phi'_i\Phi'_j\Phi'_k\right)\delta(\overline{\theta\theta}) + h.c. + \Phi'^{\dagger}_i\Phi'_i \\ &= \left(t_ie^{-iq_i\alpha}\Phi_i + \frac{1}{2}m_{ij}e^{-i(q_i+q_j)\alpha}\Phi_i\Phi_j + \frac{1}{3}\gamma_{ijk}e^{-i(q_i+q_j+q_k)\alpha}\Phi_i\Phi_j\Phi_k\right)\delta(\overline{\theta\theta}) + h.c. \\ &\quad + \Phi^{\dagger}e^{iq_i\alpha}e^{-iq_i\alpha}\Phi\end{aligned} \tag{4.149}$$

The kinetic energy part of the Lagrangian is invariant automatically, but the potential part is invariant only under conditions

$$\begin{aligned} t_i &= 0 \quad \text{if } q_i \neq 0 \\ m_{ij} &= 0 \quad \text{if } q_i + q_j \neq 0 \\ \gamma_{ijk} &= 0 \quad \text{if } q_i + q_j + q_k \neq 0 \end{aligned} \tag{4.150}$$

That is, the potential term has to be charge-neutral if one wants to maintain the (general) charge conservation.

4.6.2
Local *U*(1) Transformation

If we want a local $U(1)$ transformation, that is, $\alpha = \alpha(x)$, the transformed chiral field is no longer a superfield because

$$\begin{aligned}\overline{D}_{\dot\alpha}\left(e^{-ig\alpha(x)}\Phi\right) &= -ig\left(\overline{D}_{\dot\alpha}\alpha(x)\right)e^{-ig\alpha(x)}\Phi + e^{-ig\alpha(x)}\overline{D}_{\dot\alpha}\Phi \\ &\neq e^{-ig\alpha(x)}\overline{D}_{\dot\alpha}\Phi\end{aligned} \tag{4.151}$$

Thus, in order for $e^{-iq_i\alpha(x)}\Phi$ to be a chiral superfield, we must require that $\overline{D}_{\dot{\alpha}}\alpha(x)$ vanishes, which means that the phase $\alpha(x)$ itself has to be a chiral superfield, that is, $\alpha(x) = \Lambda(x, \theta, \overline{\theta})$, which satisfies the condition $\overline{D}_{\dot{\alpha}}\Lambda(x, \theta, \overline{\theta}) = 0$. But then the kinetic part of the Lagrangian transforms as

$$\Phi'^{\dagger}\Phi' = \Phi^{\dagger} e^{i\Lambda^{\dagger}} e^{-i\Lambda}\Phi \neq \Phi^{\dagger}\Phi \tag{4.152}$$

It follows then that the gauge invariance is maintained by introducing the gauge vector superfield that couples to Φ such that the interaction is expressed as

$$\mathscr{L}_{\Phi} = \Phi^{\dagger} e^{2gV}\Phi \;^{7)} \tag{4.153}$$

The transformed vector superfield [Eq. (4.107)]

$$V'(x, \theta, \overline{\theta}) \equiv V(x, \theta, \overline{\theta}) + \frac{i}{2g}[\Lambda(x, \theta, \overline{\theta}) - \Lambda^{\dagger}(x, \theta, \overline{\theta})] \tag{4.154}$$

compensates the phase change of the matter field. Substituting Eqs. (4.90) and (4.111) in Eq. (4.153), the explicit expression for Eq. (4.153) is given by [411]

$$\begin{aligned}\int d^2\theta d^2\overline{\theta}\, \Phi^{\dagger} e^{2gV}\Phi &= \int d^2\theta d^2\overline{\theta}\left[\Phi^{\dagger}\Phi + 2g\Phi^{\dagger}V\Phi + 2g^2\Phi^{\dagger}V^2\Phi\right] \\ &= |F|^2 + i\partial_{\mu}\overline{\psi\sigma}^{\mu}\psi + \partial_{\mu}A^*\partial^{\mu}A + igV^{\mu}\left[A^*\partial_{\mu}A - A\partial_{\mu}A^*\right] \\ &\quad - gV_{\mu}\overline{\psi\sigma}^{\mu}\psi - \sqrt{2}g\left[\overline{\lambda}\,\overline{\psi}A + \lambda\psi A^*\right] + g\left[D(x) + gV_{\mu}V^{\mu}\right]|A|^2 + \text{total derivatives}\end{aligned} \tag{4.155}$$

As is customary for the gauge interaction, the above expression can be reformulated into a compact form using the covariant derivative

$$D_{\mu} = \partial_{\mu} - igV_{\mu} \tag{4.156}$$

Then, Eq. (4.155) becomes

$$\begin{aligned}\int d^2\theta d^2\overline{\theta}\, \Phi^{\dagger} e^{2gV}\Phi &= |F(x)|^2 + i\left(D^*_{\mu}\overline{\psi}(x)\right)\overline{\sigma}^{\mu}\psi(x) + D^*_{\mu}A^*(x)D^{\mu}A(x) \\ &\quad - \sqrt{2}g\left[\overline{\lambda}(x)\,\overline{\psi}(x)A(x) + \lambda(x)\psi(x)A^*(x)\right] + gD(x)|A(x)|^2\end{aligned} \tag{4.157}$$

Notice that the Yukawa coupling is not arbitrary as in the SM, but is given by the same coupling constant as the gauge interaction. Then, the $U(1)$ gauge-invariant supersymmetric action is given by

$$\begin{aligned}S = \int d^4x d^2\theta d^2\overline{\theta}\Big[&\Big\{W^{\alpha}W_{\alpha}\delta(\overline{\theta\theta}) + \overline{W}_{\dot{\alpha}}\overline{W}^{\dot{\alpha}}\delta(\theta\theta)\Big\} \\ &+ \Phi^{\dagger}e^{2gV}\Phi + \Big\{W[\Phi]\delta(\overline{\theta\theta}) + h.c.\Big\}\Big]\end{aligned} \tag{4.158}$$

where the kinetic terms of the vector superfields in the first line are given by Eq. (4.147) and $W[\Phi]$ is the superpotential.

---

7) The factor 2 looks odd but it is a matter of the vector field normalization. It is introduced to make the covariant derivative take the conventional form given in Eq. (4.156). If we do not attach the factor 2, the covariant derivative becomes $D_{\mu} = \partial_{\mu} - igV_{\mu}/2$.

As there is no derivatives of the D field in the Lagrangian, it also is an auxiliary field that can be re-expressed in terms of other fields. The kinetic part contains $D^2(x)/2$, and the gauge interaction contains $gD(x)\,|A|^2$. Combining them, one can construct another potential consisting of the field $D(x)$. Therefore, we obtain

$$\begin{aligned} &\tfrac{\partial\mathcal{L}}{\partial D} = D(x) + g|A|^2 = 0, \quad \rightarrow \quad D(x) = -g|A|^2 \\ &V_D \equiv -\tfrac{1}{2}D^2(x) - gD(x)|A|^2 = +\tfrac{1}{2}\left|gA^*A\right|^2 \end{aligned} \tag{4.159}$$

where V_D is the counter part to V_F in the chiral field Lagrangian and is commonly referred to as the *D-term*. The total potential is given by sum of F- and D-terms.

$$V = V_F + V_D = |F(x)|^2 + \frac{1}{2}|D(x)|^2 \tag{4.160}$$

4.6.3
Non-Abelian Interaction

For the non-Abelian case, the vector field, the D-term, and so on are modified to

$$W_\alpha^{\mathrm{NA}} = \frac{1}{4g}\overline{DD}e^{-gV}D_\alpha e^{gV}, \qquad \overline{W}_{\dot\alpha}^{\mathrm{NA}} = \frac{1}{4g}DDe^{-gV}\overline{D}_{\dot\alpha}e^{gV} \tag{4.161a}$$

$$D^{\mathrm{NA}\,a}(x) = -g\sum_{ij} A_i^*[T^a]_{ij}A_j \tag{4.161b}$$

$$V_D^{\mathrm{NA}} = \frac{1}{2}\sum_a\sum_{ij}\left|gA_i^*[T^a]_{ij}A_j\right|^2 \tag{4.161c}$$

$$[D_\mu]_{ij}^{\mathrm{NA}} = \delta_{ij}\partial_\mu - ig\sum_a V_\mu^a[T^a]_{ij} \tag{4.161d}$$

where $[T^a]_{ij}$ is the generators of the non-Abelian symmetry group. Equation (4.161a) reduces to Eq. (4.112) for the Abelian case.

For the chiral part, all the derivatives in the chiral Lagrangian Eq. (4.142) have to be replaced with covariant derivatives. Furthermore, the chiral field must belong to a representation of the gauge group. As SUSY and gauge transformations commute, the scalar, fermion, and auxiliary fields must be in the same representation, so

$$\delta_{\mathrm{gauge}}X_i = ig\sum_a \Lambda_a(T^aX)_i \tag{4.162}$$

for $X_i = A_i, \psi_i, F_i$. The superpotential must satisfy the condition

$$\delta_{\mathrm{gauge}}W = \sum_i \frac{\partial W}{\partial A_i}(T^aA)_i = 0 \tag{4.163}$$

For example, requirement of $SU(2)$ invariance constrains a pair of the scalar fields to be mutually conjugate, and one of the scalar fields in the Yukawa potential has to be a gauge singlet.

4.7
Summary of SUSY Lagrangian

Starting from definitions of the following chiral superfields Φ_L, Φ_R and a vector superfield V_{WZ} in the Wess–Zumino gauge,

$$\begin{aligned}
\Phi_L(x,\theta,\bar\theta) &= \hat\Phi\left(y,\theta\right) = A(y) + \sqrt{2}\theta\psi(y) + \theta\theta F(y) \\
\Phi_R(x,\theta,\bar\theta) &= \hat\Phi\left(z,\bar\theta\right) = A^*(z) + \sqrt{2}\bar\theta\bar\psi(z) + \bar\theta\bar\theta F^*(z) \\
y &= x - i(\theta^\alpha\sigma^\mu_{\alpha\dot\alpha}\bar\theta^{\dot\alpha}), \quad z = x + i(\theta^\alpha\sigma^\mu_{\alpha\dot\alpha}\bar\theta^{\dot\alpha}) \\
V_{WZ}(x,\theta,\bar\theta) &= (\theta\sigma^\mu\bar\theta)V_\mu(x) + \theta\theta\bar\theta\bar\lambda(x) + \bar\theta\bar\theta\theta\lambda(x) + \theta\theta\bar\theta\bar\theta\frac{1}{2}D(x)
\end{aligned} \tag{4.164}$$

we give the full Lagrangian describing their interactions. The kinetic energy of the gauge superfield can be obtained by substituting Eq. (4.161) into Eq. (4.147) and that of the chiral superfied from Eq. (4.157). The SUSY-invariant Lagrangian containing the chiral superfield (A_i, ψ_i) and the gauge superfield (λ_a, V_a^μ) in the off-shell representation is expressed as

$$\mathcal{L}_{\text{susy}} = \mathcal{L}_{\text{gauge KE}} + \mathcal{L}_{\text{chiral KE}} + \mathcal{L}_{\text{superpotential}} \tag{4.165a}$$

$$\mathcal{L}_{\text{gauge, KE}} = \sum_a \left[\frac{1}{2}|D^a(x)|^2 - \frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} - i\lambda^a\sigma^\mu\partial_\mu\bar\lambda^a\right] \tag{4.165b}$$

$$\begin{aligned}
\mathcal{L}_{\text{chiral, KE}} &= \sum_i \left[|F_i(x)|^2 + i(D^*_\mu\bar\psi_i(x))\bar\sigma^\mu\psi_i(x) + D^*_\mu A^*(x)D^\mu A(x)\right] \\
&+ \sum_a \left[-\left\{\sqrt{2}g\bar\lambda^a(x)\left(\bar\psi(x)T^aA(x)\right) + (c.c.)\right\} + gD^a(x)\left(A^*(x)T^aA(x)\right)\right]
\end{aligned} \tag{4.165c}$$

$$[D_\mu]_{ij} = \delta_{ij}\partial_\mu - ig\sum_a V^a_\mu[T^a]_{ij}$$

$$\mathcal{L}_{\text{superpotential}} = \sum_i W_iF_i - \frac{1}{2}\sum_{j,k} W_{jk}\psi_j\psi_k + h.c.$$

$$W_i = \frac{\partial W(A)}{\partial A_i}, \quad W_{ij} = \frac{\partial^2 W(A)}{\partial A_i\partial A_j}, \quad F_i = -W_i^* = -\left[\frac{\partial W(A)}{\partial A_i}\right]^*$$

$$W(A) = t_iA_i + \frac{1}{2}m_{ij}A_iA_j + \frac{1}{3}y_{ijk}A_iA_jA_k \tag{4.165d}$$

where T^a is the gauge group generators and i, j, k denote the flavor of the chiral fields.

In the on-shell representation

$$\begin{aligned}
\mathcal{L}_{\text{susy}} = \sum_a &\left[-\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} - i\lambda^a\sigma^\mu\partial_\mu\lambda^a\right] + \left[\bar\psi(x)\bar\sigma^\mu iD_\mu\psi(x) + D^*_\mu A^*(x)D^\mu A(x)\right] \\
&- \left[\sum_{ijk}\left\{\frac{1}{2}m_{ij}\psi_i(x)\psi_j(x) + y_{ijk}\psi_i(x)\psi_j(x)A_k(x)\right\}\right. \\
&\left. + \sqrt{2}g\sum_a \lambda^a(x)\left(A^*(x)T^a\psi(x)\right) + (c.c.)\right] - (V_D + V_F)
\end{aligned} \tag{4.166a}$$

$$V_D = \frac{1}{2}\sum_a |D^a(x)|^2 = \frac{1}{2}\sum_a \sum_{ij} \left| g A_i^* [T^a]_{ij} A_j \right|^2 \tag{4.166b}$$

$$\begin{aligned} V_F &= \sum_i |F_i|^2 = \sum_i \left| \frac{\partial W(A)}{\partial A_i} \right|^2 \\ &= \sum_{ijk} \left| t_i + m_{ij} A_j(x) + 2y_{ijk} A_j(x) A_k(x) \right|^2 \end{aligned} \tag{4.166c}$$

$$\begin{aligned} &\overset{\text{if } t_i=0}{=\!=} \sum_{i,j,\, n\neq ij} m_{in}^* m_{nj} A_i^* A_j + \sum_{i,j,k,\, n\neq ijk} \left(m_{in}^* y_{jkn} A_i^* A_j A_k + m_{in} y_{jkn}^* A_i A_j^* A_k^* \right) \\ &\qquad + \sum_{i,j,k,l,\, n\neq ij,kl} y_{ijn}^* y_{kln} A_i^* A_j^* A_k A_l \end{aligned} \tag{4.166d}$$

where we added an explicit expression of V_F for the case $t_i = 0$ for later use.

We reiterate some notable features of the supersymmetric Lagrangian:

1) The superpotential $W(A)$ is an analytic function of $A(x)$ treated as a complex number. That is, it does not contain the complex conjugate fields $A^*(x)$ or derivatives of the field.
2) The potential term is sum $V_F + V_D$, which is semipositive definite.

4.8
Spontaneous Symmetry Breaking

Phenomenologically, breakdown of the SUSY is put in by hand, as we will see in the next chapter. This is because there are no universally accepted dynamic symmetry-breaking methods for the SUSY. As the potential is given by

$$V = V_F + V_D = \sum_i \left| \frac{\partial W(A)}{\partial A_i} \right|^2 + \frac{1}{2}\sum_l g_l^2 \sum_a \left| \sum_{ij} A_i^* [T_l^a]_{ij} A_j \right|^2 \tag{4.167}$$

we can break SUSY if either $< F_i > = < \partial W/\partial A_i > \neq 0$ for some i [F-term breaking] or if $D_{l,a} = g_l < \sum_{ij} A_i^* [T_l^a]_{ij} A_j > \neq 0$ for some combination of (l, a) [D-term breaking]. An example of F-term breaking is the O'Raifeartaigh model and the D-term breaking Fayet–Iliopoulos model [414, 415]. Here, we discuss a few aspects of the spontaneous symmetry breakdown in SUSY and have a glimpse of its difficulty.

A special feature of the SUSY is that the Hamiltonian and the SUSY operator Q_α are closely related.

$$\begin{aligned} H = P^0 &\overset{(4.55\text{d})}{=\!=} \tfrac{1}{4}\{Q_\alpha, \overline{Q}_{\dot\beta}\}(\overline{\sigma}^{0\,\dot\beta\alpha}) = \tfrac{1}{4}(Q_\alpha \overline{Q}_{\dot\beta} + \overline{Q}_{\dot\beta} Q_\alpha)(\overline{\sigma}^{0\,\dot\beta\alpha}) \\ &= \tfrac{1}{4}\left(Q_1 \overline{Q}_{\dot 1} + \overline{Q}_{\dot 1} Q_1 + Q_2 \overline{Q}_{\dot 2} + \overline{Q}_{\dot 2} Q_2 \right) \end{aligned} \tag{4.168}$$

Equation (4.168) means that the Hamiltonian is semipositive, that is

$$<\eta|H|\eta> \quad \geq \quad 0 \tag{4.169}$$

where $|\eta>$ is any state including the vacuum state. The vacuum $|0>$ is the lowest energy state and hence has zero energy. Now we apply the SUSY transformation to the vacuum:

$$\delta_S|0>= (\alpha^{\alpha} Q_{\alpha} + \overline{\alpha}_{\dot\alpha} \overline{Q}^{\dot\alpha})|0> \tag{4.170}$$

If the SUSY is spontaneously broken, that is, if the vacuum breaks the SUSY, we must have

$$\delta_S|0>\neq 0 \tag{4.171}$$

Then, by virtue of Eqs. (4.168) and (4.169), the Hamiltonian must also have finite energy. This feature is in contrast to other symmetries in which the vacuum energy can be negative or zero when it breaks the symmetry.

4.8.1
D-Term Breaking

SUSY breaking with a nonzero D-term VEV can occur through the Fayet–Iliopoulos mechanism [415]. One introduces a linear term κD into the Lagrangian. If the gauge symmetry is $U(1)$, it also respects the SUSY [see Eq. (4.110)]. Then, referring to Eq. (4.159), the relevant part of the off-shell Lagrangian that contains the D-term has the form [410]

$$V_D = \kappa D - gD \sum_i q_i |A_i|^2 - \frac{1}{2} D^2 \tag{4.172}$$

where q_i is the $U(1)$ charge of the scal field A_i in the chiral super-multiplet. The presence of the Fayet–Iliopoulos term modifies the equation of motion Eq. (4.159) to

$$D = \kappa - g \sum q_i |A_i|^2 \tag{4.173}$$

If the superpotential made of the chiral fields contain a nonzero mass term $\sim m_{ij} A_i A_j$, the gauge invariance forces $q_i + q_j = 0$. The corresponding potential $\sum |F_i|^2 = |\partial W(A)/\partial A_i|^2$ (see Eq. (4.166c)) contains the mass term and the potential will have the form

$$V = \sum_i |m_i|^2 |A_i|^2 + \frac{1}{2}(\kappa - g \sum_i q_i |A_i|^2)^2 \tag{4.174}$$

This potential is positive and finite; hence, the SUSY must be broken. One can see that the potential minimum occurs for nonzero D. For the simplest case in which $|m_i|^2 > g q_i \kappa$ for each i, the minimum is realized for all $A_i = 0$ and $D = \kappa$. The gaugino remains massless because it is the Goldstone particle (Goldstino) that has appeared as a consequence of spontaneous symmetry breakdown. For non-Abelian gauge groups, the analog of $\mathcal{L} = -\kappa D$ would not be gauge-invariant, and therefore is not allowed, so only $U(1)$ D-terms can drive spontaneous symmetry breaking.

The SM contains the $U(1)$ gauge symmetry; however, squarks and sleptons cannot make a superpotential mass term with zero hypercharge, and hence the minimum supersymmetric extension of the Standard Model (MSSM) cannot break the SUSY spontaneously by the D-term breakdown mechanism. It also has difficulty in producing an appropriate mass for gauginos. As the D-term breaking leads to an unacceptable mass spectrum, the alternative F-term breaking is the preferred model of the spontaneous symmetry breaking.

4.8.2
F-Term Breaking

For the F-term, O'Raifeartaigh [414] demonstrated that at least three different scalar superfields are necessary to break the symmetry spontaneously. One chooses the superpotential to be

$$W(\Phi) = g\Phi_1(\Phi_3^2 - m^2) + \lambda\Phi_2\Phi_3 \tag{4.175}$$

where λ, g, m are three real nonzero parameters. Then,

$$V_F = |F_1|^2 + |F_2|^2 + |F_3|^2 \tag{4.176a}$$

$$F_1 = \frac{\partial W(A)}{\partial A_1} = g(A_3^2 - m^2) \tag{4.176b}$$

$$F_2 = \frac{\partial W(A)}{\partial A_2} = \lambda A_3 \tag{4.176c}$$

$$F_3 = \frac{\partial W(A)}{\partial A_3} = \lambda A_2 + 2gA_1A_3 \tag{4.176d}$$

It is easy to demonstrate that one cannot choose parameters λ, g, m to satisfy the three equations $\partial W/\partial A_1 = \partial W/\partial A_2 = \partial W/\partial A_2 = 0$ simultaneously. Therefore, the potential is nonzero, and hence the symmetry is spontaneously broken. As one of the F_is gets the VEV, it is referred to as *F-term breaking*. The potential is given by

$$V_F = g^2m^4 + (\lambda^2 - 2g^2m^2)A_3^2 + g^2A_3^4 + |\lambda A_2 + 2gA_1A_3|^2 \tag{4.177}$$

If $\lambda^2 - 2g^2m^2 > 0$, which we will assume, the absolute minimum of the potential is at $A_2 = A_3 = 0$, with A_2 taking any value. It is referred to as the *flat direction* in the potential if one or more of the fields can have any value. One can show that one of the three fermions (ψ_j's) in the chiral superfield is massless, which is the Goldstino.

Proof: The condition for the potential minimum is given by

$$\frac{\partial V_F}{\partial A_i} = \frac{\partial}{\partial A_i}\sum_j \left|\frac{\partial W}{\partial A_j}\right|^2 = \sum_j (W_{ij}^* W_j + W_j^* W_{ij}) = 0$$

$$W_{ij} = \frac{\partial^2 W(A_1, A_2, A_3)}{\partial A_i \partial A_j} \tag{4.178}$$

Therefore, at the potential minimum

$$W_{ij} W_j^* = 0 \tag{4.179}$$

As W_{ij} is the fermionic mass matrix [see Eq. (4.165d)], and $W_j \neq 0$ for some j, the equation means that at least one of the eigenmasses vanishes. ■

Notice that the O'Raifeartaigh model again contains the linear term in the superpotential, which has to be a gauge singlet. We do not have such candidate in the MSSM. Besides, it is generally hard to give masses to the gauginos because the SUSY invariant Lagrangian does not have any (scalar–gaugino–gaugino) couplings that could turn into gaugino mass when the scalar gets the VEV. It can also be shown from general grounds that, in the SUSY-invariant formalism, at least some of the MSSM squarks and sleptons would have to be unacceptably light and should have been discovered already [410].

The scale of the symmetry breaking is set by $\sqrt{F}$. If one wants the SUSY to solve the big hierarchy problem, the natural scale for the symmetry breaking is the GUT (Grand Unified Theory) energy; yet, it has to be of the electroweak scale to control the radiative corrections to the Higgs mass. For these reasons, one expects that the soft symmetry breaking in MSSM arises indirectly or radiatively, rather than tree-level couplings to the spontaneous SUSY breaking. Evidently, the SUSY breaking occurs in a "hidden sector" that has no direct couplings to the visible sector. This will be discussed in the next chapter when we build the supersymmetric models.

5
Supersymmetry II: Phenomenology

5.1
Introduction

Despite the absence of evidence for the supersymmetry (SUSY), physicists like it because of its many attractive features. It can solve the hierarchy problem, achieve gauge unification, and provide a likely candidate for the dark matter (DM). The superstring theory, the leading theory to unite all the forces, is constructed on it. Although a low-energy realization of the SUSY at the teraelectronvolt scale is not a necessary consequence, it has been a leading contender for a new physics beyond the SM.

SUSY means that the fermion and the boson are not different but the same particles in a different state. If the SUSY exists in Nature, there must exist a supersymmetric partner to every known particle, a fermion to a boson and a boson to a fermion. We may call the partner a supersymmetric particle or superparticle or simply ‘‘sparticle.’’

Particles and their partners (sparticles) belong to the same multiplet referred to as super-multiplet with identical mass; namely, the supersymmetric partners share the same quantum numbers with the SM particles but their spin differs by half a unit. They interact with other particles with the same coupling strength as the SM counterpart.

The stated conditions were necessary to solve the fine-tuning problem in the GUTs. In reality, no degenerate fermion–boson pairs exist; hence the SUSY must be broken. But if the breaking is small, that is, if the mass difference relative to the known SM particles is less than $O(1\,\text{TeV})$, it still provides a viable solution to the hierarchy problem. Considering the fact that the gauged SUSY leads to gravity (referred to as the supergravity (SUGRA)), model constructions pursuant to the SUSY have a potentiality to elucidate ingredients of Nature buried deep beyond the fine-tuning problem. Below we consider rules and phenomena that result from the SUSY.[1]

---

[1] The contents of this chapter owe much to [410]. Other references are [405, 406, 416, 417].

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

5.2
Minimum Supersymmetric Standard Model

5.2.1
Particle Spectrum

As the square of the SUSY operator vanishes ($Q_{\alpha i}^2 = 0$, $i = 1, \ldots, N$, where i denotes internal degrees of freedom), the algebra for $N = 1$ closes within a doublet with helicity (h) differing by 1/2. If a doublet $(h, h + 1/2)$ exists, antidoublet $(-h, -h - 1/2)$ must also exist by the CPT theorem. If $N = 2$, $Q_{\alpha 1} Q_{\alpha 2} \neq 0$, particles with helicity differing by 1/2 and 1 can coexist within a multiplet. If the graviton, a spin = 2 gauge particle of the gravity, exists, spin of the fundamental particles varies from 0 to 2. If one wants all the fundamental particles in one multiplet, it has to include all of $-2 \leq h \leq 2$ members and $N = 8$ is needed. Here, for simplicity, we restrict our discussions to $N = 1$ SUSY.

There is also a phenomenological reason to choose $N = 1$ SUSY. The extended ($N \geq 2$) SUSY in four-dimensional field theories does not allow for chiral fermions or parity violation as observed in the SM.

For $N = 1$, we have spin 1/2 fermions and their partners [chiral multiplet (1/2,0)], gauge bosons and their partners [gauge multiplets (1,1/2)], Higgs and its partner [a part of chiral multiplet (1/2,0)]. As an example, we list members of the SM and its model independent and minimum supersymmetric extension (referred to as *MSSM, minimum supersymmetric standard model* [408, 418, 419]) in Tables 5.1 and 5.2.[2]

The MSSM is the standard SUSY framework by which phenomenological analyses are carried out. Its Lagrangian consists of minimum form of the kinetic energy of chiral fields (also referred to as the *Kähler potential*), the gauge kinetic energy and the superpotential [see Eq. (4.165)]. The superpotential is at most cubic in the chiral superfields and contains the mass and the Yukawa interaction terms. To break the SUSY, a soft-SUSY-breaking term is included. The soft breaking means that it does not induce quadratically divergent radiative corrections and that their energy scales remain at the EW scale. As there is no universally accepted mechanism to break the SUSY spontaneously, it is provided as an external perturbation put in by hand and imposed on otherwise strict SUSY in order to construct phenomenologically viable models.

The Lagrangian is then constructed by including all possible supersymmetric interaction terms that satisfy $SU(3) \times SU(2) \times U(1)$ gauge invariance. Within the above framework, no constraints on the strength of the coupling constants are imposed other than those dictated by the SUSY.

A few comments are in order.

---

[2] To differentiate superpartners of the SM particles having the same quantum number, we attach a hat (ˆ) to the chiral superfield, a tilde (˜) to the sparticle. We also use generation index to differentiate flavors, for instance, $Q_i = (u_i\, d_i)$; $i = 1, 2, 3$ means $Q_1 = (u,\, d)$, $Q_2 = (c,\, s)$, $Q_3 = (t,\, b)$, and so on.

Table 5.1 List of chiral super-multiplets.

Notation	Spin 0	Spin 1/2	$(SU(3)_C, SU(2)_L, Y)$	Name
Q^a	$(\tilde{u}_L, \tilde{d}_L)$	(u_L, d_L)	(3 , 2, +1/3)	Squark
$\overline{u}$	$\tilde{u}_R^*$	$u_R^\dagger$	(3*, 1, −4/3)	Squark
$\overline{d}$	$\tilde{d}_R^*$	$d_R^\dagger$	(3*, 1, +2/3)	Squark
L	$(\tilde{\nu}_L, \tilde{e}_L)$	(ν_L, e_L)	(1, 2, −1)	Slepton
$\overline{e}$	$\tilde{e}_R^*$	$e_R^\dagger$	(1, 1, +2)	Slepton
H_u	$(H_u^+, H_u^0)^b$	$(\tilde{H}_u^+, \tilde{H}_u^0)$	(1, 2, +1)	Higgsino
H_d	(H_d^0, H_d^-)	$(\tilde{H}_d^0, \tilde{H}_d^-)$	(1, 2, −1)	Higgsino
		$\tilde{G}_{1/2}$		Goldstino [c]

(a) Three generations exist for the quarks and leptons. The fourth column denotes dimensions of the symmetry group and hypercharge Y of the SM.
(b) After spontaneous symmetry breakdown, the physical Higgses and their super partners are denoted as h, H, A $H^\pm$ and $\tilde{h}$, $\tilde{H}$, $\tilde{A}$, $\tilde{H}^\pm$. They further mix with gauginos to become $\tilde{\chi}_i^0$ and $\tilde{\chi}_i{}^\pm$.
(c) Goldstino appears in the global SUSY breaking, but in SUGRA, it is eaten by the gravitino denoted as $\tilde{G}$.

Table 5.2 List of vector supermultiplets.

Spin 1/2	Spin 1	$(SU(3)_C,\ SU(2)_L,\ U(1)_Y)$	Name of sparticles
$\tilde{B}^a$	B	(1, 1, 0)	Bino
$\tilde{W}^{\pm b}$	$W^\pm$	(1, 3, 0)	Wino
$\tilde{\gamma}^c$	γ		Photino
$\tilde{Z}^c$	Z		Zino
$\tilde{g}$	g	(8, 1, 0)	Gluino

(a) Bino, neutral wino, and higgsinos mix to make four neutralinos $\tilde{\chi}_1^0 \sim \tilde{\chi}_4^0$.
(b) Winos and charged higgsinos mix to make four charginos $\tilde{\chi}_1^\pm$, $\tilde{\chi}_2^\pm$.
(c) Photino and zino are special combination of neutralinos.

1) All the fermions in the SM carry chirality L or R which transform differently under the symmetry group $SU(2) \times U(1)$, that is, they carry different quantum numbers. Therefore, they have to belong to different chiral supermultiplets $(\tilde{f}_L, f_L)$ and $(\tilde{f}_R, f_R)$, respectively. The superpartners $\tilde{f}_L$ and $\tilde{f}_R$ are scalar bosons and have no chirality but denoted with suffixes L, R to distinguish their partners. They mix; that is, the mass eigenstates $(\tilde{f}_1, \tilde{f}_2)$ are mixture of them.
2) In the SM, there is only one Higgs doublet. If SUSY is required, one needs at least two of them. The mass generating Lagrangian, i.e., the Yukawa interaction

is incorporated in the superpotential. It has the form

$$-\mathscr{L}_{\text{Higgs}} = \frac{\sqrt{2}m_u}{v_u}\overline{u}_R Q_L \cdot H_u - \frac{\sqrt{2}m_d}{v_d}\overline{d}_R Q_L \cdot H_d - \frac{\sqrt{2}m_e}{v_d}\overline{l}_R L_L \cdot H_d + (h.c.) \tag{5.1a}$$

$$H_u = \begin{bmatrix} H_u^+ \\ H_u^0 \end{bmatrix}, \quad H_d = \begin{bmatrix} H_d^0 \\ H_d^- \end{bmatrix}, \quad Q_L = \begin{bmatrix} Q_L^1 \\ Q_L^2 \end{bmatrix} = \begin{bmatrix} u_L \\ d_L \end{bmatrix}, \quad L_L = \begin{bmatrix} L_L^1 \\ L_L^2 \end{bmatrix} = \begin{bmatrix} \nu_L \\ e_L^- \end{bmatrix} \tag{5.1b}$$

where $Q_L \cdot H_u \equiv Q_L^1 H_u^2 - Q_L^2 H_u^1 = u_L H_u^0 - d_L H_u^+$, and so on, denotes an $SU(2)$ invariant combination.[3] The minus sign in the second and third term is to unify the sign of the mass term when the Higgs acquire vacuum expectation value (VEV). H_u couples only to up-type quarks and H_d couples only to down-type quarks and charged leptons. In the MSSM, there are no right-handed neutrinos.

The reason that the SUSY Lagrangian needs two Higgs doublets comes from analyticity of the superpotential. If there is only one Higgs H_d as in the SM, H_u in Eq. (5.1a) has to be replaced with H_d^c. Since it is complex conjugate of H_d, it is not allowed in the superpotential (see footnote[6] in Chapter 4). Consequently, H_u belonging to a different multiplet which couples only to the up-type quarks has to be introduced. However, note that $v_u^2 + v_d^2 = 4m_W^2/g_W^2 \simeq (246\,\text{GeV})^2$ is fixed by the W mass and the $SU(2)$ gauge coupling. There is another reason to have two Higgs doublets. A higgsino doublet that is a partner to the SM Higgs by itself cannot make the triangle anomaly vanish (see Section 3.2.3). It is necessary to have two doublets to compensate each other.

Two Higgs doublets contain eight independent components. We already discussed in Section 1.4 that three of them are absorbed to give masses to $W^\pm$, Z, but the rest appear as physical particles $(h^0, H^0, A, H^\pm)$. Corresponding to the five Higgses, five higgsinos $(\tilde{h}^0, \tilde{H}^0, \tilde{A}, \tilde{H}^\pm)$ are needed. When the $SU(2)_L \times U(1)$ breaks to $U(1)_{\text{EM}}$, mixing among them occurs. $\tilde{W}^\pm$ and $\tilde{H}^\pm$ mix to become charginos $(\tilde{\chi}_{1,2}^\pm)$ and $\tilde{B}$, $\tilde{W}^0, \tilde{h}^0, \tilde{H}^0$ mix to become neutralinos $(\tilde{\chi}_i^0 \; i = 1 \sim 4)$. The numbering is in ascending order in terms of mass. The photino $(\tilde{\gamma})$ and Zino $(\tilde{Z})$ are special combinations of the bino $\tilde{B}$ and the wino $\tilde{W}^0$. $\tilde{A}$ is the partner of A which has CP = − and is made of imaginary part of two neutral Higgses. It does not mix with $\tilde{h}, \tilde{H}$ in the CP conserving world. Note that gluinos and neutralinos are Majorana fermions. This can be understood by counting the number of degrees of freedom.

3) When the SUSY breaks down spontaneously, Goldstone bosons associated with the SUSY appear. As the SUSY operator is a spinor, it has spin 1/2 and is referred to as the *Goldstino* (denoted as $\tilde{G}_{1/2}$). If the SUSY is a global symmetry, this is a physical particle with vanishing mass. In the supergravity (SUGRA), the symmetry is local and the Goldstino is absorbed to give mass to the spin 3/2 gravitino. However, it is not a part of MSSM.

---

[3] We use nomenclatures in this chapter to express fields to conform to two-component formalism adopted in the previous chapter.

4) The sparticles have SM quantum numbers identical to their partners except for their spin. From this fact, it is shown that one cannot find superpartners in the known particle spectrum. For instance, when only the spin is considered, q_L, q_R, l_L, l_R have a possibility of being partners to the gauge bosons, but looking at the gluon's quantum number in QCD (being a color octet) and that of $W^\pm, Z$ in $SU(2) \times U(1)$ (being an iso triplet), they cannot be partners. The quarks and leptons have the same $SU(2)$ quantum number as Higgs, but they cannot be the higgsinos, because if they are, the quark or lepton number is not conserved after the breakdown of the symmetry and contradicts the observations. Thus we conclude that sparticles are not found among the known particles and requirement of the SUSY doubles the number of elementary particles.
5) SUSY and SM particles are distinguished by a multiplicative quantum number R-parity defined by

$$R \equiv (-1)^{2S+3(B-L)} \tag{5.2}$$

where S, B, and L denote spin, baryon, and lepton number, respectively. It is a consequence of the need for $B - L$ conservation. All the known particles have R-parity $+1$, while all the sparticles have $R = -1$. If R-parity is violated, the exchange of sparticles may lead to an unacceptable fast decay of protons. It is therefore commonly assumed that R-parity is conserved. As a consequence, the sparticles can only be produced pair-wise and the lightest sparticle (LSP, lightest superparticle) must be stable. There are models that violate the R-parity, but we chose not to discuss them to avoid too much complication. LSP is usually the lightest neutralino, but in some models, gravitino or sneutrino is favored.

The LSP is the preferred candidate for the DM. Its existence is required by cosmological arguments and observations (see Chapters 9 and 10). It does not radiate photons and it does not interact strongly. Namely, it consists of stable neutral particles that are not members of the SM.

5.2.2
Interactions

If the SUSY breakdown is spontaneous, degeneracy of the masses would disappear, but the interaction and its coupling strength do not change. For instance, if the interaction between the fermion f and the gauge boson V is specified by its coupling strength g, the interaction between the gaugino $\tilde{V}$, sfermion $\tilde{f}$, and the fermion f can be derived [416, 420]. Replace $(A, \psi, \overline{\psi})$ and $(V^\mu, \lambda, \overline{\lambda})$ in Eq. (4.164) with $(\tilde{f}_a, f_a, \overline{f}_a)$ and $(V^\mu, \tilde{V}, \overline{\tilde{V}})$, substitute them into the interaction terms in $\sqrt{2}g\overline{\lambda}\overline{\psi}A$ and $D_\mu^* A D^\mu A$ ($D_\mu = \partial_\mu - igV_\mu$) [see Eq. (4.166)], one obtains

$$\begin{aligned} \mathscr{L}_{\text{INT}} = g\overline{f}_a \gamma^\mu f_a V_\mu + h.c. &\rightarrow \sqrt{2}g(\overline{f}_a \tilde{V})\tilde{f}_a + h.c. \\ &\rightarrow ig\left[(\partial_\mu \tilde{f}_a^*)\tilde{f}_a - \tilde{f}_a^* \partial_\mu \tilde{f}_a\right] V^\mu + h.c. \end{aligned} \tag{5.3}$$

where $f_a(\tilde{f}_a) = f_L(\tilde{f}_L)$ or $f_L^c(\tilde{f}_R)$. $\tilde{f}_a$ ($a = L, R$) is a spin 0 scalar but labeled with its partner's helicity. The two sfermions ($\tilde{f}_L$ and $\tilde{f}_R$) share the same quantum number

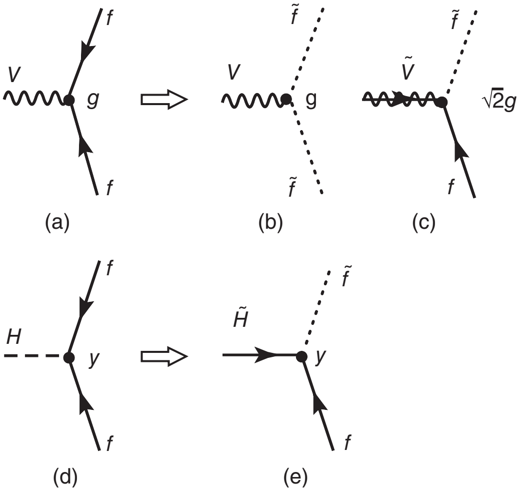


Figure 5.1 Coupling strengths of sparticles. (a) Vff in SM, (b) $V\tilde{f}\tilde{f}$, (c) $\tilde{V}f\tilde{f}$, (d) Hff in SM, and (e) $\tilde{H}f\tilde{f}$. The gaugino is denoted by wavy lines overlapped with straight lines.

and after the breakdown of EW symmetry couple to the same Higgs. As a result, they mix, and hence are not mass eigenstates [see Eq. (5.35)]. However, as the mixing rate is proportional to the mass of the fermion, mixing other than that of the third-generation fermions can be ignored. Some examples of the interaction types are depicted in Figure 5.1.

As is clear from the above argument, the form and the coupling strength of the sparticle interaction are not affected by the symmetry breakdown. Only the masses change. Consequently, if the mass spectrum of the sparticles is known, the production and decay rates can be basically predicted up to the mixing effect.

5.2.3
Constraints

In order to obtain the mass spectrum, a model is required. In order to not contradict past experiments, some constraints are imposed on the model construction. If the sparticles exist, they contribute to radiative corrections and change theoretical predictions of the observed processes (like $\mu \to e\gamma$, $K_L^0 \to \mu\overline{\mu}$, etc.), $K_L - K_S$ mass difference, CP parameters, absence of FCNC (flavor changing neutral current), and so on [421–425]. For $\tilde{u}$ and $\tilde{d}$, the atomic parity violation [see Section 2.4.1 of [2]] becomes too large unless their L and R components are almost degenerate. For others, a large mass difference between $\tilde{f}_L$ and $\tilde{f}_R$ induces intolerable FCNC effects. A recipe frequently adopted is to require degeneracy condition, which assumes almost degenerate mass spectrum for different flavors.

$$\begin{aligned}
m^2(\tilde{Q}_1) &\approx m^2(\tilde{Q}_2) \sim m^2(\tilde{Q}_3)\\
m^2(\tilde{L}_1) &\approx m^2(\tilde{L}_2) \sim m^2(\tilde{L}_3)\\
m^2(\overline{u}) &\approx m^2(\overline{c}) \sim m^2(\bar{t})\\
m^2(\overline{d}) &\approx m^2(\overline{s}) \sim m^2(\overline{b})\\
m^2(\overline{e}) &\approx m^2(\overline{\mu}) \sim m^2(\overline{\tau})
\end{aligned} \tag{5.4}$$

The constraint to the third generation can be relaxed to some extent. Another constraint comes from the CP violation. Any SUSY violating term that contains CP violating complex amplitudes should be almost real or exactly real if the CKM (Cabibbo-Kobayashi-Maskawa) mechanism is to be respected.

5.2.4
SUSY Breaking

There are more than one way to break the SUSY. However, the breaking must be soft to be applicable to the teraelectronvolt-scale phenomena. The MSSM was proposed based on the principle of simplicity and consistency with general SUSY outcomes. Its theoretical constraints are too weak and the number of parameters is as large as 124 [419, 426]. However, the phenomenological restrictions discussed in the previous section exclude much of the parameter space. The MSSM theory is viable only in the very restricted region of the parameter space.

One way to reduce the number of parameters in the MSSM is to assume a model in which one requires some kind of unification at high energy ($\sim M_X$). Typically, one requires that all the scalar particles and gauginos are unified and have common mass values m_0 and $m_{1/2}$, respectively.

However, from the model building point of view, there are no dynamically satisfactory way of breaking the SUSY spontaneously. A way to circumvent the situation is to assume that the SUSY breaking occurs in a hidden sector consisting of particles that are completely neutral with respect to the SM gauge group. Its effects are transmitted to the MSSM particles (sometimes referred to as the *visible sector*) by some messengers.

Three theoretical scenarios are popular. One is the constrained MSSM (CMSSM), traditionally known as *minimum supergravity* (mSUGRA) in which the symmetry breaking (SB) in the hidden sector is transmitted to the visible sector by the gravity. As the symmetry breaking occurs typically through the F-term by acquiring VEV, the symmetry breaking scale is of the order $<F> /M_{\text{Planck}}$. For this to be consistent with the soft breaking, $M_X \sim \sqrt{<F>} \sim 10^{11}$ GeV is required. Consequently, one naturally relies on SUGRA-GUTs for the unification scheme [427–431].

The second is general gauge mediation (GGM) in which the messenger has the quantum number of the SM $[SU(3) \times SU(2) \times U(1)]$ and interacts with the visible sector through ordinary gauge interaction [432]. Its mass must be heavy in order to not have been discovered already. In its simple version known as GMSB (gauge mediated symmetry breaking) [433–437], the messenger communicates with the hidden sector through a gauge singlet chiral multiplet (denoted as S). When the field S and its associated auxiliary field (denoted as F_S) acquire VEV, the messenger gets mass $\Lambda = <F_S> / <S>$, which, in turn, induces mass to the gaugino in the visible sector through loop corrections. The generated mass is given by $M_a \simeq (\alpha_a/4\pi)\Lambda$ ($a = 1 - 3$). If $<S> \simeq \sqrt{<F_S>}$, the value of $\Lambda \sim \sqrt{<F_S>}$ could be as small as ~100 TeV. Mediation through gravity also exists in GGM, but its effect is small as $\sqrt{<F_S>} \sim 10^{11}$ GeV is required for the gravity-mediated interaction to be sizable.

The third is the anomaly mediated symmetry breaking (AMSB). Here, the quantum anomaly of the super-Weyl-scale invariance is the source of the SB. The mechanism is universal to any SUGRA and it contributes to loop integrals giving a large mass to the gauginos. In this sense, it may be considered as a variant of the mSUGRA, although its phenomenology is considerably different from it.

A common feature of the above three scenarios is to assume some kind of unification á la GUTs reducing the number of parameters to a handful. Conventionally, masses of scalars or gauginos and the Yukawa coupling strengths are assumed to take common values at high energies.

Furthermore, there are extra-dimensional approaches, where one considers two branes (four-dimensional spacetime) in the extra dimension (ED), one assigned to the SM world with its superpartners and the other to the hidden sector. The messenger lives in the bulk, i.e., the whole D-dimensional space. The SB effect is typically reduced by a factor $e^{-M_D R}$ where R is the distance between the two branes and M_D is the fundamental Planck mass in D-dimensional spacetime. An example is a variant of AMSB mentioned above where one assumes that the symmetry breaking takes place in the second brane [438]. Another is gaugino-mediated SUSY where the gauge superfield lives in the bulk [439, 440].

These models are based on clear physical ideas and have high constraints. They contribute only to a very limited region of the parameter space in the MSSM, which may turn out to be too restrictive. Alternatively, a phenomenological approach can be taken to study a broader and more comprehensive subset of the MSSM (pMSSM, phenomenological MSSM) [418, 441, 442]. Here, experimental data are used to eliminate parameters, namely, the constraints discussed in the previous section are imposed. The constraints reduce the number of free parameters to 19, making it a practical compromise between the full MSSM and highly constrained CMSSM. In the following, we make simple discussions on CMSSM, GGM, and AMSB to have an idea of what the constrained models are like. We also refer to CNMSSM (constrained next MSSM) and split SUSY where they are appropriate. The MSSM can be obtained by relaxing the constraints imposed by those models, namely, setting the unified parameters independent of each particle species.

5.2.5
Higgs Potential

In the MSSM, the terms contributing to the scalar Higgs potential V_H come from three different sources [30].

1) The D-term containing the quartic Higgs interaction Eq. (4.166). For the two Higgs field $H_1 = H_d$ and $H_2 = H_u$ with $Y = -1$ and $+1$, these terms are given by

$$U(1)_Y : \; V_{D1} = \frac{1}{2}\left[\frac{g_B}{2}\left(|H_2|^2 - |H_1|^2\right)\right]^2$$

$$SU(2)_L : \; V_{D2} = \frac{1}{2}\sum_{a=1}^{3}\left[\frac{g_W}{2}\left(H_1^*\tau^a H_1 + H_2^*\tau^a H_2\right)\right]^2 \tag{5.5}$$

Using completeness relation, $\sum_a \tau^a_{ij}\tau^a_{kl} = 2\delta_{il}\delta_{jk} - \delta_{ij}\delta_{kl}$, one can rewrite V_{D2} and obtain

$$V_D = V_{D1} + V_{D2} = \frac{g_W^2}{8}\left[4|H_1^\dagger H_2|^2 - 2|H_1|^2|H_2|^2 + (|H_1|^2)^2 + (|H_2|^2)^2\right] + \frac{g_B^2}{8}\left(|H_2|^2 - |H_1|^2\right)^2 \tag{5.6}$$

2) The F-term of the superpotential Eq. (4.166) can be written as $V_F = \sum_i |\partial W/\partial A_i(x)|^2$. As $W = \mu H_1 \cdot H_2$ from Eq. (5.17), one obtains

$$V_F = \mu^2(|H_1|^2 + |H_2|^2) \tag{5.7}$$

3) From the SB Lagrangian that is given later in Eq. (5.18), one obtains

$$V_{\text{soft}} = m_{H_1}^2 H_1^\dagger H_1 + m_{H_2}^2 H_2^\dagger H_2 + B\mu(H_1 \cdot H_2 + h.c.) \tag{5.8}$$

where $m_{H_1}^2$, $m_{H_2}^2$ are the Higgs masses evolved from m_0^2 at the unification scale M_X down to the EW scale by renormalization group equation (RGE) [see Eqs. (3.89) and (5.15)]. The full scalar potential for the Higgs fields is the sum of (a)–(c)

$$V_H = (|\mu|^2 + m_{H_1}^2)H_1^\dagger H_1 + (|\mu|^2 + m_{H_2}^2)H_2^\dagger H_2 + B\mu(H_1 \cdot H_2 + h.c.) + \frac{g_B^2 + g_W^2}{8}\left(|H_1|^2 - |H_2|^2\right)^2 + \frac{1}{2}g_W^2|H_1^\dagger H_2|^2 \tag{5.9}$$

Expanding the Higgs fields in terms of charged and neutral components and defining the mass-squared terms

$$H_d = \begin{bmatrix} H_d^1 \\ H_d^2 \end{bmatrix} = \begin{bmatrix} \phi_1^{0\dagger} \\ -\phi_1^- \end{bmatrix}, \quad H_u = \begin{bmatrix} H_u^1 \\ H_u^2 \end{bmatrix} = \begin{bmatrix} \phi_2^+ \\ \phi_2^0 \end{bmatrix}, \tag{5.10a}$$

$$\mu_1^2 = |\mu|^2 + m_{H_1}^2, \qquad \mu_2^2 = |\mu|^2 + m_{H_2}^2, \qquad \mu_3^2 = -B\mu \tag{5.10b}$$

we can write down the MSSM Higgs potential.

$$V_H = \mu_1^2(|\phi_1^0|^2 + |\phi_1^-|^2) + \mu_2^2(|\phi_2^0|^2 + |\phi_2^+|^2) - \mu_3^2(\phi_1^-\phi_2^+ + \phi_1^{0\dagger}\phi_2^0 + h.c.) + \tfrac{g_W^2+g_B^2}{8}\left(|\phi_1^0|^2 + |\phi_1^-|^2 - |\phi_2^0|^2 - |\phi_2^+|^2\right)^2 + \tfrac{g_W^2}{2}\left|\phi_1^{0\dagger}\phi_2^+ - \phi_1^+\phi_2^0\right|^2 \tag{5.11}$$

which reproduces Eq. (1.46).

At the SUGRA scale ($\sim M_X$), we have $\mu_1^2 = \mu_2^2$ because $m_{H_1}^2 = m_{H_2}^2 = m_0^2$. It does not satisfy the condition for the spontaneous symmetry breaking. However, the relation holds only in the neighborhood of the SUGRA scale. Inclusion of radiative corrections make these parameters evolve as a function of $\tau = \ln(Q^2/M_X^2)$ with different beta functions. We require that one of the Higgs mass (squared) turns into negative at the EW scale [$Q \sim O(m_Z)$]. Necessary conditions to induce spontaneous breaking are as follows:

1) Determinant of the Higgs mass matrix becomes negative,

$$\mu_1^2\mu_2^2 - \mu_3^4 < 0 \tag{5.12a}$$

2) The potential has a lower limit.

$$\mu_1^2 + \mu_2^2 - 2|\mu_3^2| > 0 \tag{5.12b}$$

Writing down the condition for the Higgs potential to have a minimum [$\partial V_H/\partial v_i = 0$ $(i = 1, 2)$] at the EW scale, we obtain two equalities:

$$\mu_1^2 = \mu^2 + m_{H_d}^2 = \mu_3^2 \tan\beta \quad - \frac{m_Z^2}{2}\cos 2\beta \tag{5.13a}$$

$$\mu_2^2 = \mu^2 + m_{H_u}^2 = \mu_3^2 \mathrm{cotan}\beta + \frac{m_Z^2}{2}\cos 2\beta \tag{5.13b}$$

$$\tan\beta = \frac{v_2}{v_1} \tag{5.13c}$$

$$\frac{m_Z^2}{2} = \frac{\mu_1^2 - \mu_2^2\tan^2\beta}{\tan^2\beta - 1} = \frac{m_{H_d}^2 - m_{H_u}^2\tan^2\beta}{\tan^2\beta - 1} - \mu^2 \tag{5.13d}$$

The above equalities satisfy the necessary conditions Eq. (5.12). Utilizing the above formula, we can rewrite μ^2 and $\mu_3^2 = B\mu$ in terms of m_Z^2 and $\tan\beta$. In constrained models, one typically starts from a limited number of parameters. In mSUGRA, for instance, one specifies five parameters

$$(m_0,\ m_{1/2},\ A_0,\ \tan\beta,\ \mathrm{sign}[\mu]) \tag{5.14}$$

Then all the sparticle masses can be derived from the above parameters as evolution results from the RGEs. The meaning of the variable A_0 is given later in Eq. (5.18b). The parameters m_0, $m_{1/2}$, and A_0 are real if we do not consider CP violation or almost real to be consistent with observations. The master equation for calculating the mass spectrum at the EW scale is expressed as (see Eq. (1.25)(3.9))

$$\frac{dm_i^2}{d\tau} = \sum_j \beta_{ij} m_j^2(\tau), \qquad \tau = \ln\frac{Q^2}{m_X^2} \tag{5.15}$$

In order to have spontaneous breakdown of the $SU(2) \times U(1)$, either one of the Higgs mass has to go negative. If the mass of the top quark is large enough, the coupling with the Higgs H_2 is strong. Then the right-hand side of Eq. (5.15) for the H_2 mass is positive and large, making the μ_2^2 an increasing function of τ. If its value is positive at the SUGRA scale, it can go negative at the EW scale. Figure 5.2 demonstrates such an example of evolutions of the Higgs mass [443].

This is a salient feature of the SUGRA. In the SM, the EW spontaneous symmetry breakdown had to be put in by hand, whereas in SUGRA it can be derived from the symmetry breakdown at the Planck scale supplemented by the RGEs.

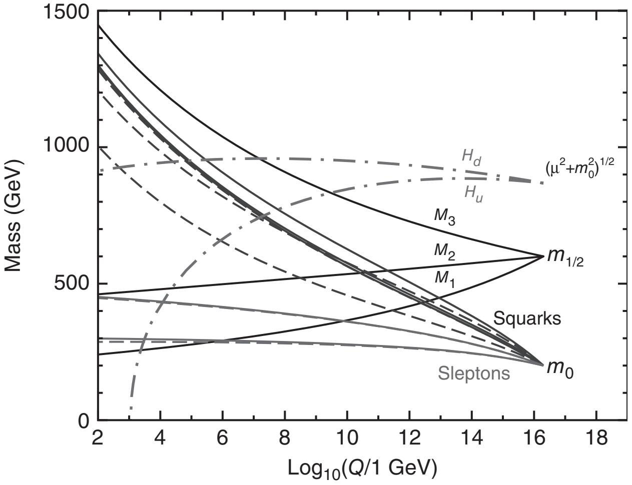


Figure 5.2 Behavior of the sparticle mass in mSUGRA as a function of Q starting from $M_X = 10^{16}$ GeV. The bold lines are the three gaugino masses $(\tilde{g}, \tilde{W}$, and $\tilde{B})$. The light solid lines are the squark masses $(\tilde{q}_L, \tilde{q}_R, \tilde{t}_L, \tilde{t}_R)$ and slepton masses$(\tilde{l}_L, \tilde{l}_R)$. The dashed lines represent the Higgs boson masses (H_d, H_u). Because of the large top mass, the mass of the H_u is driven to negative inducing the spontaneous symmetry breakdown at the electroweak scale. (Reproduced with permission of [410, 443].)

It can also reproduce the top mass approximately (100–200 GeV) [444–446].

$$m_t \simeq 196\ \text{GeV}[1 + 2\alpha_3(m_Z) - 0.12)] \sin\beta \tag{5.16}$$

Note, however, that the squared mass of the squarks and sleptons should not go negative, because then the baryon and/or lepton number is no longer conserved. Figure 5.2 also demonstrates evolution of various masses [443]. Evolution dependence on the flavor is discussed in detail in the next section.

5.3
Minimum SUGRA

5.3.1
Soft-SUSY Breaking

In models that have SUGRA incorporated within them, the process that is responsible for separation of the gravity from other forces also has an effect on the EW symmetry breaking. Suppose the superunified symmetry group (call it G including the gravity) is broken at the energy scale $M_X \simeq < F_S >$, slightly below the Planck scale M_{Planck}. $< F_S >$ is the VEV of the F field, which breaks the SUSY spontaneously. At this stage, all the coupling constants and masses are unified assuming common values except that are affected by the group structure. At energy scale $Q < M_X$, they evolve in a different manner following the RGEs.

Among those, one of the Higgs masses m_H^2 becomes negative at the EW scale $Q \sim O(G_F^{-1/2})$, inducing the EW symmetry breaking in a most natural way [447].

As stated earlier, we assume that the SB occurs in a hidden sector, which means the breaking affects only those hitherto unknown particles that do not interact with the SM particles (including their superpartners). The assumption is that those particles in the hidden sector and those in the visible sector interact only through gravity. Then even if the SB occurs at $\sim M_X$ scale, the strength of the symmetry breaking force is effectively reduced by some powers of the Planck mass. In order for the effect to be of the order of EW strength, that is, $(M_X^2/M_{\text{Planck}}) \sim O(G_F^{-1/2})$, the symmetry breaking scale can be set as $M_X \sim 10^{11}$ GeV. The resultant symmetry structure is almost equivalent to that of global SUSY GUTs and reproduces low-energy results phenomenologically equivalent to MSSM. For this reason, mSUGRA is also referred as the constrained MSSM (CMSSM). But the SUGRA serves as a useful guide in many model buildings. The gravitino denoted as $\tilde{G}$ (or $\tilde{g}_{3/2}$ in the old literature) has a mass of the EW scale, but its interaction being only gravitational there are no phenomenologically detectable effects.

In the following we first discuss only the mSUGRA with simplest assumptions. We can identify chiral superfields that include SM particles. They are as follows: $SU(2)$ doublets $\hat{Q}_i^a$, $\hat{L}_i^a$, $\hat{H}_1^a$, $\hat{H}_2^a$ ($a = 1, 2$: isospin up and down) and $SU(2)$ singlets $\overline{u}_i, \overline{d}_i, \overline{e}_i$. The index i denotes generation number. Then the most general superpotential compatible with gauge invariance, renormalizability, and R-parity conservation is written as [see Eq. (4.165)]

$$W = \sum_{ij=1}^{3} \left\{ y_u^{ij} \overline{u}_i \hat{Q}_j \cdot \hat{H}_u - y_d^{ij} \overline{d}_i \hat{Q}_j \cdot \hat{H}_d - y_l^{ij} \overline{l}_i \hat{L}_j \cdot \hat{H}_d \right\} + \mu \hat{H}_u \cdot \hat{H}_d \tag{5.17}$$

$y_{u,d,e}^{ij}$ denote Yukawa couplings among generations. The first three terms in Eq. (5.17) are SUSY generalization of the Yukawa interaction in the SM, while the last term is a globally SUSY Higgs mass term. Using Eq. (5.17), one can write down the SUSY interaction Lagrangian.

The effective Lagrangian can be expressed as a sum of SUSY invariant and SB Lagrangian:

$$\mathscr{L} = \mathscr{L}(\text{SUSY}) + \mathscr{L}_{\text{SB}} \tag{5.18a}$$

$$-\mathscr{L}_{\text{SB}} = m_0^2 \sum_{i=\tilde{q},\tilde{l},H_{1,2}} |\phi_i|^2 + m_{1/2} \sum_{a=\tilde{B},\tilde{W},\tilde{g}} \overline{\lambda}_a \lambda_a + B\mu H_u \cdot H_d + A_0 W_3 + h.c. \tag{5.18b}$$

$$W_3 = \sum_{ij=1}^{3} \left\{ y_u^{ij} \tilde{\overline{u}}_i \tilde{Q}_j \cdot H_u - y_d^{ij} \tilde{\overline{d}}_i \tilde{Q}_j \cdot H_d - y_l^{ij} \tilde{\overline{l}}_i \tilde{L}_j \cdot H_d \right\} \tag{5.18c}$$

The first and the second terms in Eq. (5.18b) are mass terms of the scalar fields and the gaugino fields. Attachment of mass only to the superpartners is the cause for the SB. The third and fourth terms look like the SUSY conserving superpotential in Eq. (5.17), but here instead of using the superfields with hat (^) on top, only superpartners of the SM particles (denoted with tilde (~) are picked up and hence they also break the SUSY.

In this form, SB is soft and does not induce quadratic divergences [394, 448]. All the constants are of the teraelectronvolt energy scale. Remember that all the SM particles are massless before the breaking of EW symmetry, which is conserved at this stage. A_0 and B are parameters with mass dimension. m_0 and $m_{1/2}$ are the common masses of the chiral scalars and gauginos and $A_0 y_k^{ij}$ is the common Yukawa coupling at M_X scale.

$$m^2(\tilde{q}) = m^2(\tilde{l}) = m^2(H_u) = m^2(H_d) = m_0^2 \tag{5.19a}$$

$$M(\tilde{g}) = M(\tilde{W}^{\pm}) = M(\tilde{B}) = m_{1/2} \tag{5.19b}$$

$$A(\tilde{q}) = A(\tilde{l}) = A_0 \tag{5.19c}$$

Notice that common mass settings at M_X scale for all the sfermions imposes degeneracy condition Eq. (5.4) required by observations. In principle, all the SB parameters $(m_{1/2}, m_0, A_0, B\mu)$ can be complex, although they are generally assumed to be real to avoid CP violation in the SUSY sector. They develop mass differences as they evolve in going down the energy scale, but corrections remain controlled as will be shown later. As the symmetry breaking mass scale m_0, $m_{1/2}$ should not exceed $\sim O(1)$ TeV from a phenomenological point of view, A_0 and B should also be of the same order. In the MSSM, these parameters differ from particles to particles and the number of independent parameters is high [426]. In mSUGRA, governed by the renormalization equation, they differ at low energies but are unified at M_X scale increasing its predictive power enormously. At low energies when the EW symmetry breaks at $\lesssim O(1)$ TeV, particles with the same quantum number mix.

Our next task is to connect the mSUGRA assumptions in Eq. (5.19) at $\sim M_X$ scale and phenomenology of the SUSY particle spectra at the EW scale.

5.3.2
Mass Formula

Gaugino As $\beta_{ij} \propto \alpha_i$ for the gauginos, their mass evolves in the same way as the corresponding coupling constants [405, 406, 410, 416, 417] . The evolution equations of the coupling constants $\alpha_i(Q^2)$ were given in Eq. (3.89). Accordingly, one can derive the following relations:

$$M_i = \frac{\alpha_i(Q^2)}{\alpha_i(M_X^2)} m_{1/2} \tag{5.20a}$$

$$M_3 : M_2 : M_1 = g_s^2(\tau) : g_W^2(\tau) : \frac{5}{3} g_B^2(\tau) \tag{5.20b}$$

$$\frac{g_B^2}{g_W^2} = \tan^2 \hat{\theta}_W \tag{5.20c}$$

Here, $M_3 = M(\tilde{g})$, $M_2 = M(\tilde{W})$, $M_1 = M(\tilde{B})$ and the factor 5/3 comes from the GUT normalization of g_B [see Eq (3.33)]. At the EW scale,

$$M_3 : M_2 : M_1 \approx 6 : 2 : 1 \tag{5.21}$$

which means that the gluino is considerably heavier compared to the wino $(\tilde{W})$ or bino $(\tilde{B})$.

The wino consists of charged $\tilde{W}^{\pm}$ and neutral $\tilde{W}^0$. When the $SU(2) \times U(1)$ symmetry is broken, the charged gaugino $\tilde{W}^{\pm}$ and the charged higgsino $\tilde{H}^{\pm}$ mix to make two kinds of charginos ($\tilde{\chi}_1^{\pm}$, $\tilde{\chi}_2^{\pm}$).

$$\mathscr{M}(\tilde{\chi}^{\pm}) = \begin{bmatrix} M_2 & \sqrt{2} m_W \sin\beta \\ \sqrt{2} m_W \cos\beta & \mu \end{bmatrix} \tag{5.22}$$

where the off-diagonal element arises from the gauge coupling of the chiral superfields, that is, between wino and higgsino [see the last line of Eq. (4.166a)].

$$g_W \tilde{W}^+ H_u^0 \tilde{H}_d^- \sim (g_W v_u/\sqrt{2}) \tilde{W}^+ \tilde{H}_d^- = (\sqrt{2} m_W \sin\beta) \tilde{W}^+ \tilde{H}_d^- \tag{5.23}$$

Notice that the mass matrix $\mathscr{M}(\tilde{\chi}^{\pm})$ is not symmetric. Therefore, one needs two different diagonalization matrices for the negatively charged and positively charged left-handed components of the charginos. However, the product of $\mathscr{M}$ and $\mathscr{M}^T$ is symmetric. Therefore, the mass-squared eigenvalues can be obtained by diagonalizing $\mathscr{M}(\tilde{\chi}^{\pm})\mathscr{M}(\tilde{\chi}^{\pm})^T$.

$$\begin{aligned} m^2(\tilde{\chi}_{1,2}^{\pm}) &= \frac{1}{2}\Big[(M_2^2 + \mu^2 + 2m_W^2) \\ &\quad \mp \left\{(M_2^2 + \mu^2 + 2m_W^2)^2 - 4(\mu M_2 - m_W^2 \sin 2\beta)^2\right\}^{\frac{1}{2}}\Big] \end{aligned} \tag{5.24}$$

For $|\mu \pm M_2| \gg m_W$, the lighter chargino is wino like ($\tilde{\chi}_1^{\pm} \sim \tilde{W}^{\pm}$), and the heavier one is higgsino-like ($\tilde{\chi}_2^{\pm} \sim \tilde{H}^{\pm}$) with masses [410]

$$m(\tilde{\chi}_1^{\pm}) \simeq M_2 - \frac{m_W^2(M_2 + \mu \sin 2\beta)}{\mu^2 - M_2^2} \tag{5.25a}$$

$$m(\tilde{\chi}_2^{\pm}) \simeq |\mu| + \frac{m_W^2(\mu + M_2 \sin 2\beta)}{\mu^2 - M_2^2} \tag{5.25b}$$

Similarly, neutral members, the bino ($\tilde{B}$), the wino ($\tilde{W}^0$), and the higgsinos ($\tilde{h}^0$ and $\tilde{H}^0$) mix to constitute four neutralinos. One has to diagonalize the mass matrix of the flavor eigenstates $\lambda^0 = (\tilde{B}, \tilde{W}, \tilde{H}_d^0, \tilde{H}_u^0)$ to obtain masses of the neutralinos $m_{\tilde{\chi}_i^0}$, $i = 1-4$. The numbering is in ascending order in the mass. In general, these states form four distinct Majorana fermions because the gauginos belong to adjoint representation of the gauge symmetry. They are eigenstates of the symmetric mass matrix:

$$\mathscr{M}(\tilde{\chi}^0) = \begin{bmatrix} M_1 & 0 & -c_\beta s_W m_Z & s_\beta s_W m_Z \\ 0 & M_2 & c_\beta c_W m_Z & -s_\beta c_W m_Z \\ -c_\beta s_W m_Z & c_\beta c_W m_Z & 0 & -\mu \\ s_\beta s_W m_Z & -s_\beta c_W m_Z & -\mu & 0 \end{bmatrix} \tag{5.26}$$
$$c_\beta = \cos\beta, \quad s_\beta = \sin\beta, \quad s_W = \sin\hat{\theta}_W, \quad c_W = \cos\hat{\theta}_W$$

Exact formulas for the solutions are given in [449–451]. As the expressions are complicated, we only list approximate expressions valid for $|\mu \pm M_{1,2}| \gg m_Z$.

$$m(\tilde{\chi}_a^0 : a = 1,2) \simeq M_a - \frac{m_Z^2(s_W^2, c_W^2)(M_a + \mu \sin 2\beta)}{\mu^2 - M_a^2} \tag{5.27}$$

Some examples of special cases are as follows:

$$\begin{array}{lll} \text{if } M_1,\ M_2 \ll m_Z,\ |\mu| & \rightarrow & \tilde{\chi}_1^0 \sim \tilde{\gamma} \\ \text{if } M_1,\ m_Z \ll M_2,\ |\mu| & \rightarrow & \tilde{\chi}_1^0 \sim \tilde{B} \\ \text{if } M_2,\ m_Z \ll M_1,\ |\mu| & \rightarrow & (\tilde{\chi}_1^0,\ \tilde{\chi}_1^\pm) \sim (\tilde{W}^0,\ \tilde{W}^\pm) \end{array} \tag{5.28a}$$

Conversely,

$$\text{if } |\mu|,\ m_Z \ll M_1, M_2 \ \rightarrow\ \tilde{\chi}_1^0,\ \tilde{\chi}_2^0,\ \tilde{\chi}_1^\pm \sim \tilde{h}^0,\ \tilde{H}^0,\ \tilde{H}^\pm \tag{5.28b}$$

Squarks and Sleptons Masses of the squarks and sleptons can be obtained by writing down the RGEs [Eq. (5.15)] explicitly. Neglecting small coefficients, those of the first and second generations can be approximated by the following formula [452–454]. Defining $\tilde{Q}_{1L} = (\tilde{u}_L, \tilde{d}_L)$, $\tilde{Q}_{2L} = (\tilde{c}_L, \tilde{s}_L)$, $\tilde{L}_{1L} = (\tilde{\nu}_{eL}, \tilde{e}_L)$, and $\tilde{L}_{2L} = (\tilde{\nu}_{\mu L}, \tilde{\mu}_L)$, we have.[4)]

$$\tilde{m}^2(\tilde{Q}_{1L}) = \tilde{m}^2(\tilde{Q}_{2L}) = m_0^2 + K_3 + K_2 + \frac{1}{36}K_1 + \Delta(\tilde{Q}_L) \tag{5.29a}$$

$$\tilde{m}^2(\tilde{u}_R) = \tilde{m}^2(\tilde{c}_R) = m_0^2 + K_3 \qquad + \frac{4}{9}K_1 + \Delta(\tilde{u}_R) \tag{5.29b}$$

$$\tilde{m}^2(\tilde{d}_R) = \tilde{m}^2(\tilde{s}_R) = m_0^2 + K_3 \qquad + \frac{1}{9}K_1 + \Delta(\tilde{d}_R) \tag{5.29c}$$

$$\tilde{m}^2(\tilde{L}_{1L}) = \tilde{m}^2(\tilde{L}_{2L}) = m_0^2 \qquad + K_2 + \frac{1}{4}K_1 + \Delta(\tilde{L}_L) \tag{5.29d}$$

$$\tilde{m}^2(\tilde{e}_R) = \tilde{m}^2(\tilde{\mu}_R) = m_0^2 \qquad + \quad K_1 + \Delta(\tilde{l}_R) \tag{5.29e}$$

where K_1, K_2, and K_3 are loop contributions of the gauginos in $U(1)$, $SU(2)$, $SU(3)$, respectively, and are given by

$$K_a\,(a = 1 \sim 3) = \begin{bmatrix} \frac{3}{5} \\ \frac{3}{4} \\ \frac{4}{3} \end{bmatrix} \times \frac{1}{\pi} \int_{\ln Q^2}^{\ln Q_0^2} d\tau\, \alpha_a |M_a(\tau)|^2 \tag{5.30}$$

Q_0 is the starting scale of the RGEs. Coefficients of K_1 in Eq. (5.29) are values of the hypercharge $(Y/2)^2$. Δs are hyperfine structure splitting accompanying the EW symmetry breakdown and are expressed as

$$\Delta(\tilde{f}_L) = (I_3 - Q_{\text{EM}} \sin^2 \hat{\theta}_W) m_Z^2 \cos 2\beta + m_f^2 \tag{5.31a}$$

$$\Delta(\tilde{f}_R) = \qquad Q_{\text{EM}} \sin^2 \hat{\theta}_W m_Z^2 \cos 2\beta + m_f^2 \tag{5.31b}$$

Their effect resolves degeneracy due to $SU(1) \times U(1)$. For instance, separation of $\tilde{u}_L$ from $\tilde{d}_L$ can be written down explicitly as

$$\Delta(\tilde{u}_L) = \left[\ \ 1/2 - (2/3) \sin^2 \hat{\theta}_W\right] m_Z^2 \cos 2\beta + m_u^2 \tag{5.32a}$$

$$\Delta(\tilde{d}_L) = \left[-1/2 + (1/3) \sin^2 \hat{\theta}_W\right] m_Z^2 \cos 2\beta + m_d^2 \tag{5.32b}$$

[4)] Note that there are no right-handed sneutrinos ($\tilde{\nu}_R$) in the MSSM.

m_f of the first and second generations are negligible. For numerical evaluations, we set $Q_0 = M_X \approx 2 \times 10^{12}$ GeV and obtain values of K_a at $Q \sim 1$ TeV.

$$K_1 \approx 0.15\, m_{1/2}^2, \quad K_2 \approx 0.5\, m_{1/2}^2, \quad K_3 \approx (5.8 \pm 1.0)\, m_{1/2}^2 \tag{5.33}$$

Namely, we have $K_3 \gg K_2 \gg K_1$. From these values we see that important factors to determine the masses of the squarks and sleptons are m_0 and $m_{1/2}$.

Looking into more detail, we see that all the squarks of the first and second generations are almost degenerate. But the squarks are generally heavier than sleptons. This is due to the large K_3 correction involving the gluino. Without K_3 the sleptons have masses very close to m_0. We also see that $m(\tilde{f}_L) > m(\tilde{f}_R)$ because the $SU(2)$ gauginos (K_2 contribution) discriminate one from the other. If m_0 is very small, all the scalar particles are light. Within multiplets, the hyperfine structure term $\Delta(\tilde{f}_{L,R})$ resolves the $SU(2) \times U(1)$ degeneracy, generally giving $m(\tilde{l}_L)^2 = m(\tilde{\nu}_l)^2 - m_W^2 \cos 2\beta$. For $\tan\beta > 1$[5] or $\cos 2\beta < 0$, which is phenomenologically preferred, $m(\tilde{l}_L) > m(\tilde{\nu}_l)$. An important relation is derived from the interplay of parameters. From Eqs. (5.20) and (5.29), one can deduce

$$m_{\tilde{q}} \geq 0.85 m_{\tilde{g}} \tag{5.34}$$

This is largely due to the large contribution of the gluino in the squark mass.

L–R Mixing Among the Sfermions So far we have neglected the left–right mixing effect. The mixing is induced by the Yukawa couplings mediated by the Higgs and its strength is proportional to the mass; hence, it can be neglected for the first and second generations. Mass eigenstates of top-squarks (stop) $(\tilde{t}_1, \tilde{t}_2)$ can be obtained by diagonalizing the mass matrix expressed in $(\tilde{t}_L, \tilde{t}_R)$ base [455, 456].

$$\begin{aligned}
\tilde{m}^2(\tilde{t}) &= \begin{bmatrix} \tilde{m}^2(\tilde{t}_L), & m_t(A_t - \mu^* \cot\beta) \\ m_t(A_t - \mu \cot\beta), & \tilde{m}^2(\tilde{t}_R) \end{bmatrix} \\
\tilde{m}^2(\tilde{b}) &= \begin{bmatrix} \tilde{m}^2(\tilde{b}_L), & m_b(A_b - \mu^* \tan\beta) \\ m_b(A_b - \mu \tan\beta), & \tilde{m}^2(\tilde{b}_R) \end{bmatrix}
\end{aligned} \tag{5.35}$$

Expressions for $\tilde{m}^2(\tilde{f}_{L,R}), f = t, b$ can be obtained from $\tilde{m}^2(\tilde{Q}_{L,R}), f = u, d$ in Eq. (5.29) by replacing $m_u \to m_t$, $m_d \to m_b$. The off-diagonal part of the first matrix comes from SUSY invariant contribution $H_d \tilde{t}_L \tilde{t}_R^* \sim v_1 \mu y_t$ [the second term in Eq. (4.166d)] and SB contribution $H_u \tilde{t}_L \tilde{t}_R^* \sim v_2 A_t y_t$ (the first term in Eq. (5.18c)). Using $m_t = y_t v_2$, $v_1 = v \cos\beta, v_2 = v \sin\beta$, one gets the mass matrix in Eq. (5.35). $A_{t,b}$ are the A_0 parameter of Eq. (5.18b) for the top or bottom quark evolved down to the EW scale. By mixing, the mass splitting between the left- and right-handed stop and sbottom occurs, simultaneously pushing their absolute values down. Especially,

$$\tilde{m}(\tilde{t}_1) \ll \tilde{m}(\tilde{t}_2) \lesssim \tilde{m}(\tilde{Q}_{1,2}) \tag{5.36}$$

and probably $\tilde{m}(\tilde{b}_1) \ll \tilde{m}(\tilde{b}_2)$, $\tilde{m}(\tilde{\tau}_1) \ll \tilde{m}(\tilde{\tau}_2)$ also hold (needs $\tan\beta \gg 1$). Consequently, $\tilde{t}_1$ and $\tilde{\tau}_1$ are most likely the lightest squark and slepton.

---

[5] Naively, $\tan\beta = \dfrac{v_2}{v_1} \sim \dfrac{y_b}{y_t}\dfrac{m_t}{m_b} \simeq \dfrac{m_t}{m_b} \simeq 35$ is expected.

The derived mass formulas satisfy the condition Eq. (5.4) and do not induce large FCNC. The origin of the near degeneracy can be traced back to the common mass at $Q \sim M_X$, an essential feature of the model.

Qualitative features of what we have described are already depicted in Figure 5.2, which is often used as a benchmark spectrum for analyzing experimental data.

5.3.3
μ Problem

The supersymmetric Higgs potential Eq. (5.11) contains the mass parameter μ in addition to those coming from the soft symmetry breaking Lagrangian Eq. (5.18b). All the parameters in the soft breaking terms are of the EW scale. However, the parameter μ appears in the superpotential Eq. (5.17) as $\mu H_u H_d$, which is naturally of the GUT scale. To ensure that H_u and H_d get a nonzero VEV after the EW symmetry breaking, μ should be of the EW scale, too. This creates another hierarchy problem referred to as the *μ problem*.

A cure for this problem is to add a singlet Higgs superfield in addition to the two Higgs doublets in the MSSM and is referred to as the *nMSSM (next to MSSM)* [457]. It induces an extra CP-even Higgs and CP-odd Higgs, as well as the fifth neutralino referred to as the *singlino*. Two terms are added to the MSSM Lagrangian

$$W = \lambda S H_u H_d + \frac{\kappa}{3} S^3 + \cdots \tag{5.37}$$

which replaces the usual $\mu H_u H_d$ term in MSSM. When the singlet Higgs acquires a VEV $<S>$, the first term gives $\lambda <S> H_u H_d = \mu_{\text{eff}} H_u H_d$ solving the μ problem. In addition to λ and κ, other parameters added in NMSSM are the mass of the singlet Higgs m_S, two Yukawa couplings A_κ, A_λ.

One can construct a CNMSSM model in line with the CMSSM(mSUGRA) [458, 459]. CNMSSM imposes the unification of the scalar mass m_0, the gaugino mass $m_{1/2}$, and the Yukawa couplings A_0 at the GUT scale. κ can be expressed in terms of m_Z. Consequently, CNMSSM depends on the four parameters $m_{1/2}$, m_0, A_0, and λ and can be even more restrictive than CMSSM.

5.4
GMSB

GMSB (gauge-mediated symmetry breaking) [433–436, 460] is another soft SB model and is capable of producing results similar to that of the mSUGRA[6] Just like in the mSUGRA, the symmetry is broken in the hidden sector, but the communicator is not the gravity but messenger particles that couple with visible sector, that is, MSSM particles, through the gauge interaction in the SM.

---

[6] The GGM, a slightly generalized version of GMSB, is obtained from the minimum GMSB by relaxing some parameter relations. [432] The GGM will be mentioned when it is appropriate.

Phenomenologically, GMSB shares many features in common with mSUGRA with some exceptions, notably the LSP being most likely the very light (~eV) gravitino.

5.4.1
Messenger Particles

In GMSB, one introduces a new left-handed chiral supermultiplet $\hat{\Phi} \sim (\hat{Q}, \hat{L})$, $(\hat{Q}^c, \hat{L}^c)$ as messengers of the SB. In its simplest version, the quantum numbers are assumed to be that of $SU(5)$. Namely, $(\hat{Q}, \hat{L})$ and $(\hat{Q}^c, \hat{L}^c)$ belong to **5** and **5*** in $SU(5)$ or expressed alternatively in $(SU(3),\ SU(2),\ Y)$ as [see Eq. (3.11a)]

$$\begin{aligned} &\hat{Q}\ \left(3,2,-\tfrac{2}{3}\right), \qquad &&\hat{Q}^c \sim \left(3^*,2,\tfrac{2}{3}\right) \\ &\hat{L} \sim (1,2,1), &&\hat{L}^c \sim (1,2,-1) \end{aligned} \tag{5.38}$$

Furthermore, one introduces another chiral superfield "$\hat{S}$", which does not interact with the known particles but couples with the messenger through a SUSY potential

$$W(\text{SUSY}) = \gamma_2 \hat{S}\hat{L}^\dagger \hat{L} + \gamma_3 \hat{S}\hat{Q}^\dagger \hat{Q} \tag{5.39}$$

γ_2, γ_3 are Yukawa coupling constants. The multiplet $\hat{S}$ is treated as a particle here, but conceptually it can be considered as a dynamical object representing the SB mechanism. When the SUSY is broken, the scalar component of $\hat{S}$ (denoted as S) and associated auxiliary field F_S acquire VEV $<S>$, $<F_S>$ and induce masses to the fermionic and scalar components of Q, L denoted as Q_f, L_f and S_Q, S_L. [436]

$$\begin{aligned} m(Q_f) &= \gamma_3 <S> \equiv M, \qquad &m(L_f) &= \gamma_2 <S> \simeq M \\ m(S_Q) &= \gamma_3 M\sqrt{1 \pm \tfrac{\Lambda}{M}}, &m(S_L) &= \gamma_2 M\sqrt{1 \pm \tfrac{\Lambda}{M}} \\ \Lambda &\equiv \tfrac{<F_S>}{<S>} \end{aligned} \tag{5.40}$$

M represents the scale of the messenger and Λ represents the mass branching associated with the SB. The messenger interacts with the SM particles via normal gauge interaction and gives masses to them, which, in turn, breaks the SUSY in the visible sector (i.e., the SM particles and their superpartners).

5.4.2
Mass Formula

Gauginos (λ_i's) acquire mass through the one-loop interaction of the messenger as shown in Figure 5.3.

There are no one-loop contributions to the scalar particles ($\tilde{q}$, $\tilde{l}$) as the exchange of both gauge and messenger particles is necessary. Two-loop contributions can give mass to the scalar particles as shown in Figure 5.4.

If $<F_S> \ll \gamma_i <S>^2$, the loop diagrams give masses to the gaugino

$$M_i = N\frac{\alpha_i}{4\pi}\Lambda = N\frac{\alpha_i}{4\pi}\frac{<F_S>}{<S>} \tag{5.41}$$

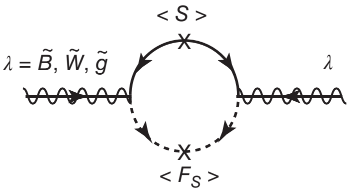


Figure 5.3 Mass acquiring mechanism of the gaugino due to the one-loop contribution of the messenger Φ. The dotted line is its scalar part and the real line is its fermionic part. The cross is the mass term associated with SUSY breaking.

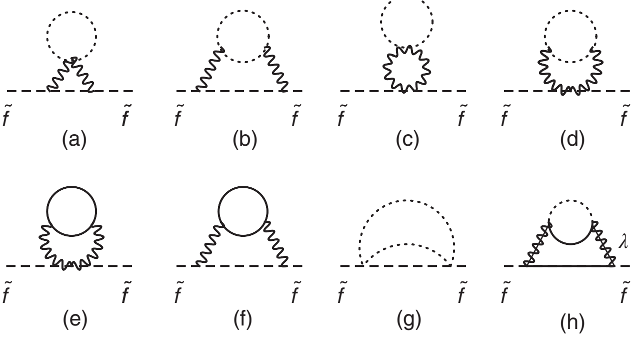


Figure 5.4 Mechanism to give mass to the sfermion ($\tilde{f} = \tilde{q}$ or $\tilde{l}$). One-loop diagrams do not contribute. Wavy lines represent the ordinary gauge particles. The fermionic and scalar part of the messenger are denoted by real and dotted lines, respectively.

and to the scalar particles

$$\tilde{m}^2 = 2\Lambda^2 N\left[C_3\left(\frac{\alpha_3}{4\pi}\right)^2 + C_2\left(\frac{\alpha_2}{4\pi}\right)^2 + \frac{3}{5}\left(\frac{Y}{2}\right)^2\left(\frac{\alpha_1}{4\pi}\right)^2\right] \tag{5.42}$$

where N is the number of messenger multiplets and in the minimum GMSB N is set to 1. $C_i\delta_{jk} = \sum_a (t^a t^a)_{jk}$ are the Casimir operators of the group and are already given as the coefficients of K_is that appeared in Eq. (5.30). $C_3 = 4/3$, 0 denote color triplets and singlets, $C_2 = 3/4$, 0 denote weak isospin doublets and singlets, and Y is the $U(1)$ hypercharge. As the SB scale is $\sim O(1)$ TeV, Eq. (5.41) means that $\Lambda \sim O(10-100)$ TeV, if no other constraints are necessary to the messenger masses. If $M \sim< S >\sim O(< F_S >^{1/2})$, the SB scale can be as small as $< F_S >\sim 10^4 – 10^5\ \text{GeV}^2$. This is in sharp contrast to the mSUGRA, where the SB scale is $M_X \sim 10^{11}$ GeV ($M_X^2 \sim< F_S >$ if the SB is due to the F-term).

Among the five soft breaking terms listed in Eq. (5.18b), contributions to A parameter is of two-loop type multiplied with α_i. Therefore, it can be neglected at the messenger scale. Instead of B parameters, one can use $\tan\beta$ as inputs as discussed in the mSUGRA section. The number of parameters in GMSB is four. They are

$$(\tan\beta,\ M,\ \Lambda,\ \text{sign}[\mu]) \tag{5.43}$$

If $N \neq 1$ the number of parameters increases to five.

5.4.3
Features of GMSB

Comparing with mSUGRA, the first salient feature of the GMSB is that the symmetry breaking scale is much smaller than that of the mSUGRA.

The second feature is that the messenger interacts with particles in the visible sector via the gauge interaction and is flavor independent. A mechanism that has to be put in by hand in mSUGRA is automatic in GMSB.

The third feature is that at the starting point, that is, at the messenger scale, the mass spectra are already split among the gaugino, squark, and sleptons. However, it is possible to consider that at the larger scale ($Q \sim M_X \sim M_{\rm GUT}$), they are unified and the difference at the messenger scale is produced by RGE evolutions starting from there. In this scenario, it is probably necessary to have $N > 1$. The gauge hierarchy in the mSUGRA as given by Eq. (5.20) is valid in GMSB, too. To go down the energy scale lower than the messenger scale, it is only necessary to further evolve the RGEs. Then the negative value of one of the Higgs VEV is also realized. In GMSB, $m_0 = 0$ has to be set in the mass formula [Eq. (5.29)] for the scalar particles. But masses already differ at the messenger scale, resulting in more or less similar spectra to those of mSUGRA.

Gravitino The fourth and probably the most salient feature of the GMSB is the existence of a light gravitino. If the SUSY is local, the massless Goldstino is eaten by the gravitino when the symmetry breaks down spontaneously. The gravitino is the spin 3/2 superpartner of the graviton, the gauge particle of the gravity. It is massless with helicity ±3/2 but acquires longitudinal components (helicity ±1/2) when it becomes massive by eating the Goldstino. It is denoted as $\tilde{G}$. The mass of the gravitino is $m(\tilde{G}) \sim <F_S> / M_{\rm Planck}$. The expression follows from a simple dimensional analysis. The gravitino mass has to vanish in the limit SUSY is restored ($<F_S> \to 0$) or the gravity is turned off ($M_{\rm Planck} \to \infty$). If $\sqrt{<F_S>}$ approaches $10^{10} \sim 10^{11}$ GeV, the mass becomes comparable to the soft SB scale, namely, it becomes as heavy as any other superparticles. Under normal circumstances, the gravitino mass is very small. A calculation shows [461].

$$M(\tilde{G}) = \frac{<F_S>}{\sqrt{3} M_{\rm Planck}} = 2.5 \left(\frac{<F_S>}{(100\ {\rm TeV})^2} \right) {\rm eV} \tag{5.44}$$

Consequently, the LSP in most GMSB models is the gravitino. If the gravitino is the LSP, it is a good candidate for the DM.

The lightest neutralino $\tilde{\chi}_1^0$, which otherwise is the primary candidate for the DM, becomes the second lightest sparticle (NLSP, next to lightest superparticle). Note, however, that depending on the model, stau ($\tilde{\tau}$) can be the NLSP. The NLSP decays to gravitino by the following reactions:

$$\tilde{\chi}_1^0 \to \gamma\,(or\ Z) + \tilde{G}, \qquad \tilde{\tau}_R^{\pm} \to \tau^{\pm} + \tilde{G} \tag{5.45}$$

If $\tilde{\chi}_1^0$ is close to $\tilde{B}$, the lifetime of the above decay modes is given by [461–463]

$$\Gamma(\tilde{\chi}_1^0 \to \gamma \tilde{G}) = \frac{\cos^2 \hat{\theta}_W \tilde{m}^5(\tilde{\chi}_1^0)}{16\pi < F_S >^2} \tag{5.46}$$

This translates to the decay length

$$c\tau \simeq 130 \left(\frac{100 \text{ GeV}}{M(\tilde{B})} \right)^{-5} \left(\frac{\sqrt{< F_S >}}{100 \text{ TeV}} \right)^4 \mu\text{m} \tag{5.47}$$

which means a finite or a very long life time for gravity-mediated interaction and can be measured. The emitted photon is "isolated" or accompanying charged particles are measured as displaced vertex or kink as far as the detection algorithm is concerned and offers a powerful method for identification of the gravitino.

One may wonder why the gravitino associated decay can be so fast. This is because the Goldstino has nongravitational interaction, which is transferred to the gravitino. If $\sqrt{< F_S >} \sim 10^{11}$ GeV, the gravitino mass is of the same order as that of other superparticles. This is also the case for the mSUGRA. With its gravitational coupling strength, it can be treated similarly to the neutrino, that is, as a noninteracting particle appearing only as missing energy in the accelerator experiments. In cosmology, however, such a heavy gravitino is problematic because it is stable and its primordial density could easily exceed the critical density. One has to think of a later inflation to dilute it or a mechanism by which it decays rapidly.

Benchmark reference points for the GMSB models are produced in [437].

5.5
AMSB and Extra Dimension

AMSB models including those residing in the ED may be considered as a variant of the SUGRA because their SB mechanism is common to all the SUGRA models. Seeking the hidden sector in ED is an attractive idea. In this case, the hidden sector of mSUGRA is another brane (a four-dimensional spacetime embedded in a $4 + 1$ dimensional hyperspace (referred to as the *bulk*), which is located at distance R apart from our world (MSSM brane) in the extra-dimensional (y) direction (see Figure 5.5). Details of the extra-dimensional space are discussed in Chapter 6. Here, we consider a simple version of one extra spatial dimension.

If a chiral superfield acquires VEV $< F_S >$ as a result of the SUSY breakdown, its strength in our world (MSSM brane) is reduced by a factor e^{RM_5} where M_5 is the five-dimensional Planck scale. Any flavor violating effect will be reduced by the exponential factor and, unlike the mSUGRA, no fine-tuning is necessary to obtain the flavor independence. If the gauge supermultiplet resides in the bulk, it becomes the messenger of the SB. Its effect on the gaugino mass is direct and is proportional to

$$M_a \sim \frac{< F_S >}{M_5(RM_5)} \tag{5.48}$$

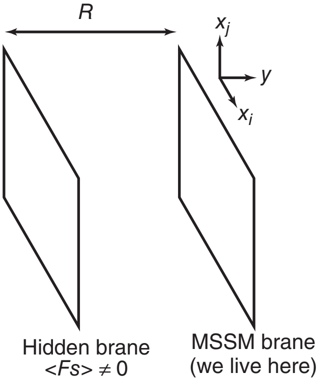


Figure 5.5 The hidden sector in SUGRA is the brane world, which is a distance R separate in the extra-dimensional (y) direction. The strength of the superfield in the hidden brane is typically reduced at MSSM brane by a factor e^{RM_5} where M_5 is the five-dimensional Planck scale.

but the effect on the scalar mass has to go through loop diagrams. Hence, the soft-SUSY breaking is dominated by the gaugino (referred to as the *gaugino mediation*). The gaugino mediation automatically reduces the flavor violation. Consequently, the above scenario is used in many models [464–466].

In AMSB [467], the symmetry breaking occurs in the hidden sector owing to quantum anomaly that breaks the Weyl-scale invariance in the supersymmetric extension of the field theory. One may assume the hidden sector to be ED [438]. Rescaling anomalies in the SUGRA Lagrangian always give rise to soft mass parameters in the observable sector. Therefore, the anomaly contributes to SB universally in all the SUGRA models. However, in many models, there are additional tree-level direct couplings that transmit the SB to the visible sector. We refer to AMSB as the model where there are no direct interactions at the tree level. The symmetry breaking masses of the gauginos are generated at one loop. As the anomaly is topological in origin, it naturally conserves flavor, thereby inducing no new FCNC amplitudes. AMSB models thus preserve virtues of the gravity-mediated models with the FCNC problem resolved. The number of parameters in the AMSB is only four, the ratio of two VEV $\tan\beta$, the gravitino mass $m_{3/2}$, the scalar mass m_0, and the sign of μ all given at the GUT scale. Low-energy gravitino mass (M_λ), scalar masses ($m_{\tilde{Q}}$), and trilinear couplings (A_y) are given by [468, 469].

$$M_\lambda = \frac{\beta_g}{g} m_{3/2} \qquad \text{gaugino mass} \qquad (5.49a)$$

$$M_{\tilde{Q}}^2 = -\frac{1}{4}\left(\frac{\partial\gamma}{\partial g}\beta_g + \frac{\partial\gamma}{\partial y}\beta_y\right) m_{3/2}^2 + m_0^2 \qquad \text{scalar mass} \qquad (5.49b)$$

$$A_y = -\frac{\beta_y}{y} m_{3/2} \qquad \text{trilinear coupling} \qquad (5.49c)$$

where β_g and β_y are beta functions of the gaugino and the trilinear Yukawa couplings, and γ is the anomalous dimension. They are functions that characterize the RGE (see Section 7.2.4 and 8.5.1 of [2]). Note that the AMSB model in its simplest version gives the tachionic mass for sleptons (i.e., negative mass-squared). The flavor-independent scalar mass term m_0^2 is an ad hoc parameter to cure it.

The AMSB has the striking predictions that:

1) the gravitino is heavy;
2) the scale-invariant one-loop gaugino mass expressions are

$$M_3 : M_2 : M_1 = -8.3 : 1 : 2.8 \tag{5.50}$$

One sees that the AMSB phenomenology based on Eq. (5.50) significantly differs from the standard phenomenology based on Eq. (5.20).

3) typical value of μ allowed by the model imply that $M_2 < M_1 < |\mu|$ which, in turn, means that the lightest chargino and the LSP, are almost pure $SU(2)$ gauginos (i.e. winos) and are almost mass degenerate:

$$m(\tilde{\chi}_1^0) \sim m(\tilde{\chi}_1^\pm) \sim M_2, \quad m(\tilde{\chi}_2^0) \sim M_1, \quad m(\tilde{\chi}_{3,4}^0) \sim m(\tilde{\chi}_2^\pm) \sim |\mu| \tag{5.51}$$

4) both the squarks and the gauginos are much heavier than the LSP. Masses of squark are also much heavier than that of the slepton, and hence rather insensitive to m_0;
5) the CP-odd A is usually much heavier than m_Z and the lighter CP-even Higgs (h^0) is almost like the SM Higgs with tightly bound upper limit of ~120 GeV [470]. Namely, a light Higgs is a firm prediction of the AMSB and the observed light Higgs at LHC already places a severe constraint on the model.

5.6
Summary of Mass Spectra

We summarize features of the mSUGRA, GMSB, and AMSB and give a bird's eye view of the mass spectrum with a caveat that exceptions exist depending on models.

1) The LSP is the neutralino $\tilde{\chi}_1^0$ in mSUGRA and the gravitino $\tilde{G}$ in GMSB.
2) The gluino is considerably heavier than charginos and neutralinos. At the EW scale the ratios are for mSUGRA and GMSB

$$M_3 : M_2 : M_1 = 6 : 2 : 1 \tag{5.52}$$

and for AMSB

$$M_3 : M_2 : M_1 = -8.3 : 1 : 2.8 \tag{5.53}$$

3) The squark in the first and the second generations is not lighter than 80% of the gluino in the mSUGRA and approximately 60% in the simplest GMSB. This is because the gluino mass contribution is sizable in the squark mass equation Eq. (5.29).

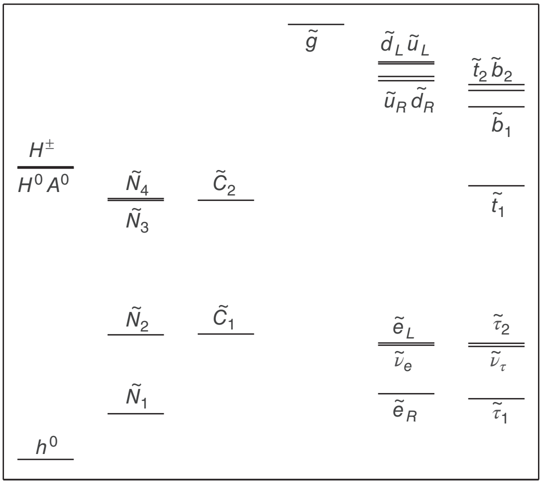


Figure 5.6 Benchmark reference mass spectrum of supersymmetric particles and Higgs bosons for the mSUGRA reference. $\tilde{N}_i, \tilde{C}_i$ are $\tilde{\chi}_i^0, \tilde{\chi}_i^\pm$ in the text. The masses of the first- and second-generation squarks and sleptons are denoted collectively by $\tilde{q}$ and $\tilde{l}$, respectively [437,471]. However, these scenarios are largely excluded by recent LHC experiments. (Reproduced with permission of [410].)

4) $\tilde{q}$ is heavier than $\tilde{l}$. Because the gluino contribution to the RGEs (K_3) is large in the evolution of the RGEs.
5) Mixing between the squarks ($\tilde{q}_L$, $\tilde{q}_R$) is small for the first and second generation because of their small mass (i.e., Yukawa coupling); hence, they are almost degenerate. However, it is large between $\tilde{t}_L$ and $\tilde{t}_R$. As a result, the mass eigenstates ($\tilde{t}_1$, $\tilde{t}_2$) are widely separated pushing the mass value themselves downward. For this reason, $\tilde{t}_1$ is considered as the lightest squark. The notion presumably applies to $\tilde{b}$ and $\tilde{\tau}$, too.
6) Generally, the left-handed sfermion [$\tilde{f}_L$ ($f = q, l$)] is slightly heavier than $\tilde{f}_R$ because the correction K_2 due to the wino is isospin dependent.
7) If the SUSY exist, the light Higgs (h^0) is lighter than 150 GeV (see Section 1.5.3, Figure 1.9). It may be even lighter than 120 GeV for AMSB. Therefore, discovery of light Higgs with mass $\sim$ 125 GeV almost excludes the simplest AMSB.

We show a representative mass spectrum of the mSUGRA in Figure 5.6 [410,471], which can serve as a benchmark reference point to facilitate collider studies.

As the LHC experiments already excluded many of the SUSY particles with mass less than $\sim$200–300 GeV, it should be regarded as a guideline in considering the overall structure of the SUSY mass spectrum.

5.7
Searches for Sparticles

We survey present situations of experimental searches for the sparticles. As the Lagrangian and the strength of the interactions are known, production cross sections and decay pattern can be predicted most of the time once their mass spectrum is assumed. In the conventional analysis, the neutralino $\tilde{\chi}_1^0$ is assumed to be the LSP. The $\tilde{\chi}_1^0$ has the EW interaction but because of *R*-parity conservation, the reaction processes contain $\tilde{q}$ or $\tilde{l}$ loops in the intermediate states and the cross section becomes proportional to $\sim\alpha^2/m(\tilde{q})^4$. As the mass of those sparticles are large, the strength of the reaction becomes of the order of weak processes or less. In other words, the LSP behaves like a heavy neutrino and rarely interacts within a detector. The signal is typically its invisibility, that is, missing energy. One can assume that the production of sparticles will always be accompanied by missing energy. In the following, we mainly discuss the so-called canonical SUSY scenario which is an mSUGRA-inspired MSSM with *R*-parity conservation (i.e., CMSSM) and has the neutralino as its LSP. We will assume the benchmark mass spectrum given in Figure 5.6 for simple consideration.

Some notable exceptions are due to GMSB. Here, the LSP is the gravitino. $\tilde{\chi}_1^0$ or possibly $\tilde{\tau}$ are the NLSP. The decay NLSP $\rightarrow \gamma\tilde{G}$ with possibly long lifetime is possible. Here, the signal for SUSY is typically a detection of missing energy accompanied by an isolated photon and/or leptons. But in GMSB the NLSP could be stable enough so that missing energy method may not be possible.

Although we discuss experimental data mainly using the CMSSM models, their constraints on the parameter space could be too tight. From phenomenological point of view, it is desirable to relax those conditions. However, from a practical point of view, the number of parameters used in the MSSM is too large. As a compromise, analyses based on pMSSM approach are often adopted. Here, the parameter spaces are restricted using the experimental data such as $K_L - K_S$ mass difference or $\mu \rightarrow e\gamma$ decay rate. Often imposed assumptions have no CP violation in the sparticle sector, minimal flavor violation, and degeneracy of the first two generations. With these restrictions, the number of parameters can be reduced to 19.

The number of parameters is still large. Because of this, simplified models are often adopted in the analysis of the hadron colliders [472–474]. It is defined by an effective Lagrangian describing the interactions of a small number of new particles. Analyses can be carried out using a small number of masses and cross sections. A typical simplified model uses one production channel and one decay channel with a 100% branching fraction (although this is not a strict requirement) in circumstances where other particles can be assumed very heavy and to decouple. Although not based on rigorous models, it serves to characterize signals of new physics if such signals are found.

5.7.1
Production Mechanism

In e^-e^+ colliders, all the sparticles, except for the gluino, are produced in a democratic way via EW interactions. If the total energy $\sqrt{s}$ exceeds twice the mass of sparticles, they are pair produced through (γ, Z) exchange in the s-channel or slepton (squark) exchange in the t-channel as illustrated in Figure 5.7. For the hadron colliders, the most copiously produced sparticles are squarks and gluinos.

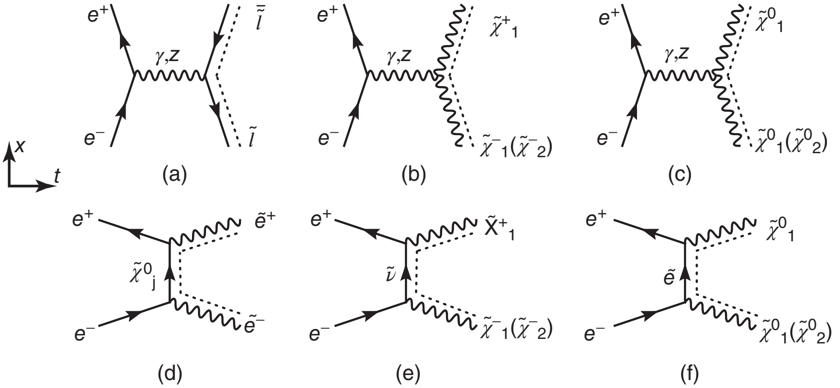


Figure 5.7 Sparticle production at e^-e^+ collider. (a)–(c) γ, Z in s-channel, (d)–(f) sparticle in t-channel. Double lines with dotted lines denote the supersymmetric particles.

As the coupling strength is known, calculation of their production cross sections and decay rates is straightforward, provided the mass and the mixing ratio are given. The production cross section is of the order 1 – 10 pb.

The fact that production and decay of Z resonance is well explained by the SM hardly allows the SUSY particles to be included in the decay products of the Z. It immediately leads to the lower limit of $m_Z/2$ for the chargino mass. The sneutrino mass greater than 43 GeV [7] is also derived. The reason that the constraint to the sneutrino mass is a bit weaker is due to smaller phase space volume contribution to the decay width.

One might think that the mass of the LSP, that is, $\tilde{\chi}^0_1$, can also be constrained. However, considering that the neutralino is a mixture of gauginos and higgsinos, its coupling strength to Z depends on the mixing ratio and is indeterminate. In the extreme case of the LSP being the photino, the coupling to Z vanishes and one cannot deduce a meaningful result from analyses of invisible width of the Z alone.

5.7.2
Sleptons

Searches by LEP In the benchmark mass spectrum of mSUGRA given in Figure 5.6, $\tilde{\mu}_R$ is lighter than $\tilde{\mu}_L$ and there is no t-channel diagram for this process. If it is the NLSP, the only decay mode is $\tilde{\mu}_R \to \tilde{\chi}^0_1 + \mu$, and the final configuration is $\mu^-\mu^+$ pair plus the missing energy that gives a distinct signal of acoplanar muon pair7) and nothing else. Figure 5.8a shows the region of $(m_{\tilde{\mu}_R} - m_{\tilde{\chi}^0_1})$ excluded by the LEPII experiments [7,475].

If the mass difference $\Delta = m_{\tilde{\mu}_R} - m_{\tilde{\chi}^0_1}$ is small ($\lesssim$5 GeV), the missing energy is small and the detection efficiency deteriorates. This is the gap between the region $m_{\tilde{\chi}^0_1} > m_{\tilde{\mu}_R}$ and the excluded region below it. The case for $\tilde{\tau}$ is similar except that

---

7) Acoplanar means the two tracks are not back-to-back but their angle is smaller than 180°.

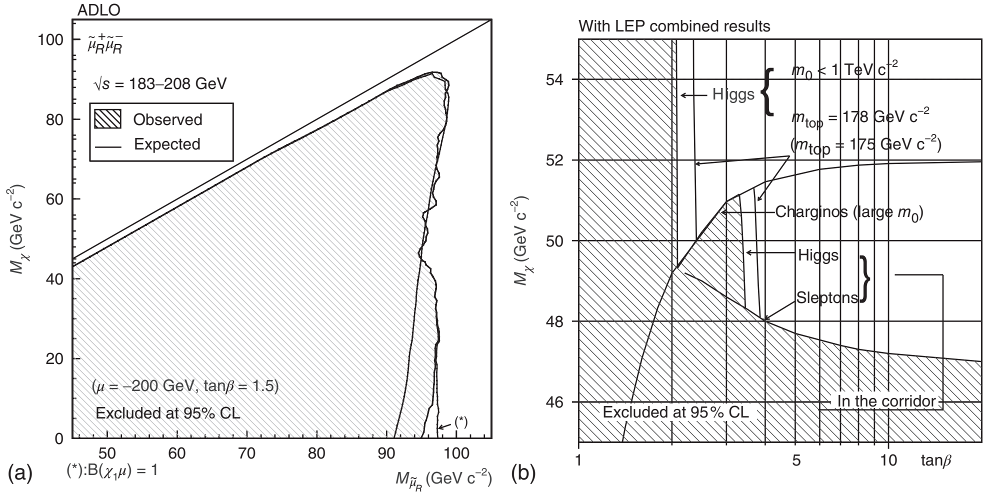


Figure 5.8 (a) Region in the $(m_{\tilde{\mu}_R}, m_{\tilde{\chi}_1^0})$ plane excluded by searches for smuons at LEP. (b) Lower mass limit for the lightest neutralino as a function of $\tan\beta$, inferred in the conventional scenario (with $M_1 \sim 0.5M_2$) from searches at LEP for charginos, sleptons, and neutral Higgs bosons. (Reproduced with permission of [7, 475].).

τ decays further to $\mu(e) + 2\nu$, but it does not change the signal except that the muon is a bit softer. A major complication is the $L - R$ mixing, which changes the coupling to Z and hence its production cross section affecting the lower limit of the $m(\tilde{\tau})$.

For the production of $\tilde{e}$, the mixing effect is negligible but additional t-channel contribution comes in (see Figure 5.7d) which enhances the cross section and as a consequence the lower limit of the mass, too. In summary, complication of mixing or t-channel contribution is small only changing the detectable mass limit slightly. One obtains [7]

$$\begin{aligned} m_{\tilde{\nu}} &> 94\ \text{GeV} \\ m_{\tilde{\mu}} &> 94\ \text{GeV} \\ m_{\tilde{\tau}} &> 81.9\ \text{GeV} \\ m_{\tilde{e}_R} &> 73\ \text{GeV} \\ m_{\tilde{e}_L} &> 107\ \text{GeV} \\ \Delta = m_{\tilde{l}} - m_{\tilde{\chi}_1^0} &\gtrsim 10\ \text{GeV} \end{aligned} \tag{5.54}$$

Searches at LHC At the hadron collider, the extraction of colorless particles (i.e., sleptons, EW gauginos, and higgsinos) is considerably harder. Many assumptions on analysis procedure and theoretical models have to be introduced [476, 477]. If the masses of the gluinos and squarks are so large as to be decoupled, the direct production of sleptons and EW gauginos $\left(\tilde{\chi}_1^\pm \tilde{\chi}_1^\pm, \tilde{\chi}_i^0 \tilde{\chi}_i^0\ (i \geq 2)\right)$ could dominate the production of SUSY particles at the hadron collider [478]. If the sleptons are heavier than the EW gauginos, they decay directly to leptons and neutralinos ($\tilde{l} \to l + \tilde{\chi}_i^0$). If the sleptons are lighter than the gauginos other than the $\tilde{\chi}_1^0$, the pair-produced EW gauginos will result in pair production of the sleptons via the following decays:

$$\begin{array}{lll} \tilde{\chi}_1^{\pm} & \to l^{\pm}\tilde{\nu},\ \ \tilde{l}^{\pm}\nu : \tilde{l}^{\pm} \to l^{\pm}\tilde{\chi}_1^0 & \text{signal:}\ \ l^{\pm} + \not{E}_T \\ & \to W^{\pm}\tilde{\chi}_1^0 & \quad ,, \\ \tilde{\chi}_{j\geq 2}^0 & \to l^{\pm}\tilde{l}^{\mp} \to l\bar{l} + \tilde{\chi}_1^0 & \text{signal:}\ \ l\bar{l} + \not{E}_T \\ & \to Z^{\pm}\tilde{\chi}_1^0 & \quad ,, \end{array} \tag{5.55}$$

Therefore, a topology including two leptons (e/μ) and the missing energies will be a good signal for the production of sleptons. Figure 5.9 shows analysis carried out by ATLAS [479] in scenarios where sleptons decay directly into the lightest neutralino and a charged lepton.

If common values are assumed for left and right-handed slepton masses, the mass between 90 and 320 GeV are excluded at 95% confidence level for a massless neutralino. In the scenario of chargino pair production, with wino-like charginos decaying into the lightest neutralino via an intermediate slepton, chargino masses between 130 and 450 GeV are excluded at 95% confidence level for a 20 GeV neutralino.

In GMSB, the LSP is the gravitino and the NLSP could have a long life. The charged NLSP is most likely the stau ($\tilde{\tau}$) and it can be searched for generically in the hadron collider. Figure 5.10 shows a result obtained by ATLAS on the search for the stau ($\tilde{\tau}$). No excess is observed above the estimated background. Long-lived

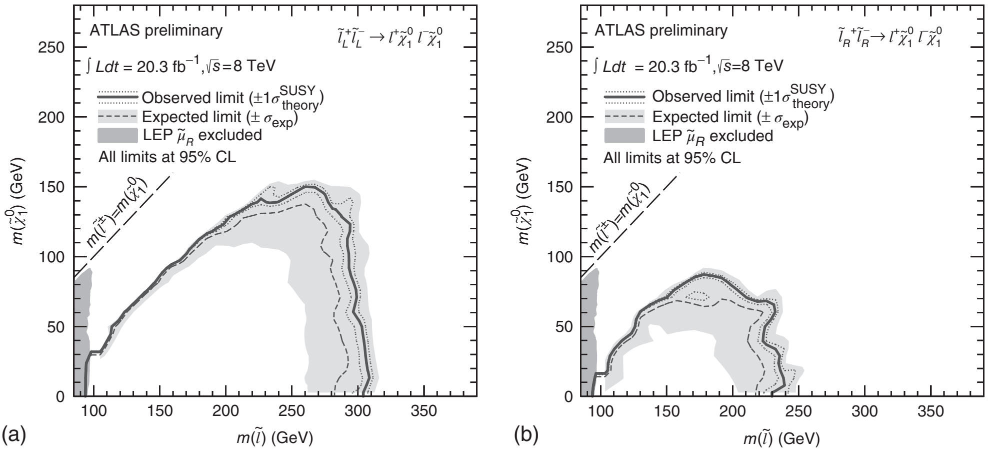


Figure 5.9 95% CL exclusion limits for (a) left- and (b) right-handed selectron and smuon production in the $m_{\tilde{l}} - m_{\tilde{\chi}_1^0}$ plane. The dashed and solid lines show the 95% CLs expected and observed limits, respectively, including all uncertainties except for the theoretical signal cross section uncertainty (PDF and scale). The solid band around the expected limit shows the $\pm 1\sigma$ result where all uncertainties, except those on the signal cross sections, are considered. Illustrated also are the LEP limits [475] on the mass of the right-handed smuon. (Reproduced with permission of [479].)

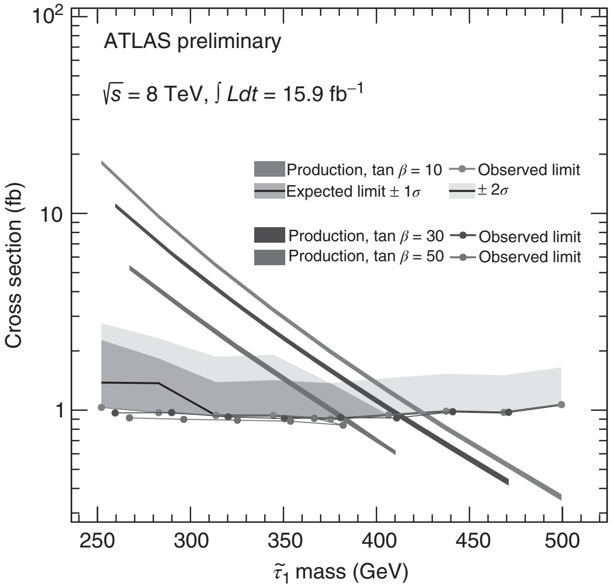


Figure 5.10 Cross section limits as a function of the $\tilde{\tau}_1$ mass in GMSB models. Observed limits are given as solid lines with markers. The theoretical prediction for the cross section is shown as a colored 1σ band. Different colors represent models with different tan β. Expected limits for tan $\beta = 10$ are drawn as magenta (top) line with ±1 and ±2σ uncertainty bands drawn in green and yellow, respectively. (Reproduced with permission of [480].) (Please find a color version of this figure on the color plates.)

$\tilde{\tau}_1$s in the GMSB model are excluded at 95% CL at masses below 385–425 GeV, for tan $\beta = 5$–50.

5.7.3
Charginos and Neutralinos

For the production of charginos and neutralinos other than the LSP, one can extract the signals using decay modes given in Eq. (5.55).

LEP results: First, we consider chargino productions. To simplify matter, let us assume that squarks and sleptons (generically denoted as $\tilde{f}$) are heavy. Then the t-channel contribution (sneutrino exchange Figure 5.7e) is negligible and the processes in consideration become $e^-e^+ \to W^+W^- + 2\tilde{\chi}_1^0$, $f\overline{f'}f''\overline{f'''} + 2\tilde{\chi}_1^0$ (e.g., $\bar{l}l\nu\overline{\nu} + 2\tilde{\chi}_1^0$), which except for the accompanying missing energy are very similar to the SM W^+W^- or $f\overline{f'}f''\overline{f'''}$ production. Hence, one can look for an excess of such processes beyond SM. No such excess were found and the limit of

$$m_{\tilde{\chi}_1^\pm} > 103 \text{ GeV for } m_{\tilde{\nu}} > 200 \text{ GeV} \tag{5.56}$$

were obtained. The limit is degraded when the mass of the sneutrino is light, partly because the t-channel interferes negatively and also because $\tilde{\chi}_1^\pm \to l\tilde{\nu}$ opens where the lepton tends to be buried in the cloud of particles and thereby reduces the efficiency for the above search mode.

The detection of neutralinos other than LSP can be pursued similarly to that of chargino. For instance, the process $e^- e^+ \to \tilde{\chi}_j^0 \tilde{\chi}_1^0 \to \tau^- \tau^+ + 2\tilde{\chi}_1^0$. Limits quoted in [7] are

$$m_{\tilde{\chi}_2^0} > 62.4\ \text{GeV}, \quad m_{\tilde{\chi}_3^0} > 99.9\ \text{GeV}, \quad m_{\tilde{\chi}_4^0} > 116\ \text{GeV} \tag{5.57}$$

However, the searches for the neutralino lose their sensitivity once the two-body decay channel $\tilde{\chi}_2^0 \to \nu\tilde{\nu}$ opens up because they are totally invisible.

LHC results: Constraints can also be obtained at the hadron collider with some assumptions. For instance, a direct production of the chargino–neutralino pair through s-channel process ($W^{\pm *} \to \tilde{\chi}_1^\pm \tilde{\chi}_2^0$), where the squark and the gluino mass were set high so that they can be neglected. Then the signal to look for is

$$pp \to W^*(Z^*) \to \tilde{\chi}^\pm(\tilde{\chi}_2^0)\tilde{\chi}_2^0 \to 3l + \not{E}_T$$

Analysis of other channels ($\tilde{\chi}_2^\pm\ \tilde{\chi}_{i\geq 3}^0$) can be done similarly but the branching ratio is typically smaller than 20%. Figure 5.11 shows analyses done by CMS [481] where the $\tilde{\chi}_1^\pm \tilde{\chi}_2^0$ are set to be wino like and mass degenerate. Two cases were analyzed, one for light sleptons (i.e., $\tilde{\chi}_1^\pm \to \tilde{l}\nu,\ l\tilde{\nu}$) and the other for heavy sleptons (i.e., $\tilde{\chi}_1^\pm \to W\tilde{\chi}_1^0$). The obtained limits are [482, 483]

$$m_{\tilde{\chi}_1^\pm} > 650\,\text{GeV} \qquad \text{for light sleptons} \tag{5.58a}$$

$$> 340\,\text{GeV} \qquad \text{for heavy sleptons} \tag{5.58b}$$

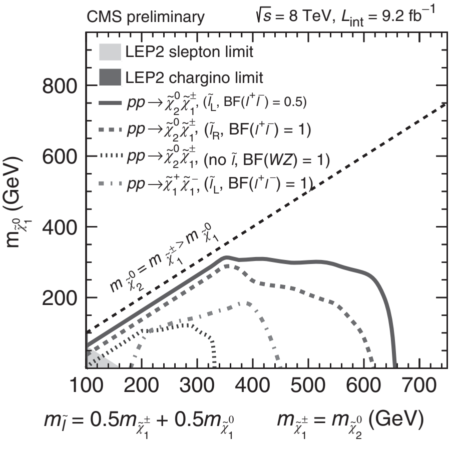


Figure 5.11 Summary of results for chargino–neutralino production with decays to left-handed sleptons, right-handed sleptons, or direct decays to vector bosons, and chargino-pair production. (Reproduced with permission of [481].)

5.7.4
LSP

Without some constraints from models, no mass limit can be obtained. Depending on the mixing ratio, LSP's coupling to Z is indeterminate and in the extreme case of $\tilde{\chi}_1^0 = \tilde{\gamma}$, it vanishes and no pair production is possible through s-channel Z exchange. The process $e^-e^+ \to \gamma\tilde{\chi}_1^0\tilde{\chi}_1^0$ is completely swamped by $e^-e^+ \to \gamma\nu\overline{\nu}$. The t-channel $\tilde{e}$ exchange has no sensitivity for heavy $\tilde{e}$.

In the mSUGRA, indirect limits can be obtained from the chargino or slepton pair production as a function of $\tan\beta$ (see Figure 5.8b). The limit obtained at LEP is

$$m_{\tilde{\chi}_1^0} > 47 \text{ GeV} \tag{5.59}$$

The absolute exclusion for $\tan\beta < 2$ is obtained from nonobservation of the light neutral Higgs boson [484].

At the LHC, the limit was increased to well above 100 GeV. However, the limit is very much dependent on models. Therefore, we will discuss the limit only in relation to searches for other SUSY particles. An example is already shown in Figure 5.11 where the mass limit of $\tilde{\chi}_1^0$ is given as a function of $m_{\tilde{\chi}_1^\pm}$. Another example will be given below in relation to the gluino production.

5.7.5
Gluino and Squarks

As the squarks and gluinos interact strongly, they can be produced copiously in the hadron collider, provided the energy is high enough. Typical processes are

$$\begin{aligned} q+\overline{q}\,(\text{or } g+g) &\to \tilde{q}\tilde{\overline{q}},\ \tilde{g}\tilde{g} \\ g+q \quad &\to \tilde{g}\tilde{q} \end{aligned} \tag{5.60}$$

If the squark is heavier than the gluino (valid in the canonical mSUGRA), the following decay chain ensues.

$$\begin{aligned} \tilde{q} \quad &\to \quad q\tilde{g} \\ \tilde{g} &\to q\overline{q}\tilde{\chi}_j^0,\ q\overline{q}'\tilde{\chi}_j^\pm \end{aligned} \tag{5.61}$$

For $j \geq 2$, further decays follow according to Eq. (5.55). If the gluino is heavier than the squarks

$$\begin{aligned} \tilde{g} &\to q\tilde{\overline{q}},\ \overline{q}\tilde{q} \\ \tilde{q} &\to q\tilde{\chi}_j^0,\ q'\tilde{\chi}_j^\pm \end{aligned} \tag{5.62}$$

The characteristic detection signals are "$n\,(\geq 1)$ jets + missing energy." The existence of isolated leptons may or may not be required depending on the decay channels one looks for. Photons in the final state can be produced via decays of higher EW gauginos. In GMSB, the neutralino decays further to the gravitino and a photon. If the neutralino is bino-like, the decay branching ratio $\tilde{\chi}_1^0 \to \tilde{G} + \gamma$ is large giving almost the same topology as the n-jet with extra one or two photons.

Experimental analyses of squark and/or gluino production typically assume the first and the second generation squarks to be approximately degenerate in mass. Production of the third generation squarks including the stop ($\tilde{q} = \tilde{t}$) will be discussed separately.

Here, we consider a simplified model [472–474] such as that described in Figure 5.12.

In the simplified model, the gluino is pair produced with subsequent decays to quark pairs and EW gauginos.

First, let us consider a final state consisting of n-jets with no leptons or photons. There are a variety of jet analysis formalism developed by the ATLAS and CMS groups. One may refer to the boxed paragraph for some of them. Figure 5.13a is a result of α_T analysis of data with a topology having two or three jets plus missing energy. One can assume the topology as a result of the direct gluino pair production with subsequent decay to a neutralino and a b-quark pair ($pp \to \tilde{g}\tilde{g} \to b\bar{b}\tilde{\chi}_1^0 b\bar{b}\tilde{\chi}_1^0$).

Several models with the gluino replaced by squarks or the final b replaced with t-quark were also tried (see Table 5.3). The result gives smaller but very similar upper limits on the mass of gluino or squarks for each process.

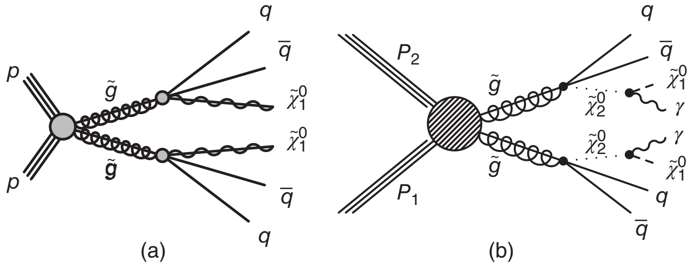


Figure 5.12 (a) An example of simplified model to produce n jets ($pp \to \tilde{g}\tilde{g}$, $\tilde{g} \to q\bar{q}\tilde{\chi}_1^0$). (b) An example in which the higher neutralino is forced to decay to a photon and the lightest neutralino. The same topology can be realized in GMSB model if $\tilde{\chi}_2^0$ is replaced with $\tilde{\chi}_1^0$ followed by subsequent decay to a photon and the gravitino.

Table 5.3 A summary of the simplified models considered in the analysis [485], which involve both direct (D) and gluino-induced (G) production of squarks, and their decays.

Model	Production/decay mode				Reference model	
					m_{parent} (GeV)	(GeV) m_{LSP}
D1	pp	$\tilde{q}\tilde{q}^*$	→	$q\tilde{\chi}_1^0\bar{q}\tilde{\chi}_1^0$	600	250
D2	pp	$\tilde{b}\tilde{b}^*$	→	$b\tilde{\chi}_1^0\bar{b}\tilde{\chi}_1^0$	500	150
D3	pp	$\tilde{t}\tilde{t}^*$	→	$t\tilde{\chi}_1^0\bar{t}\tilde{\chi}_1^0$	400	0
G1	pp	$\tilde{g}\tilde{g}$	→	$q\bar{q}\tilde{\chi}_1^0q\bar{q}\tilde{\chi}_1^0$	700	300
G2	pp	$\tilde{g}\tilde{g}$	→	$b\bar{b}\tilde{\chi}_1^0b\bar{b}\tilde{\chi}_1^0$	900	500
G3	pp	$\tilde{g}\tilde{g}$	→	$t\bar{t}\tilde{\chi}_1^0t\bar{t}\tilde{\chi}_1^0$	850	250

Models D1 and G1 concern the direct or gluino-induced production of first- or second-generation squarks only. Reference models are also defined in terms of the parent (gluino or squark) and LSP sparticles masses.

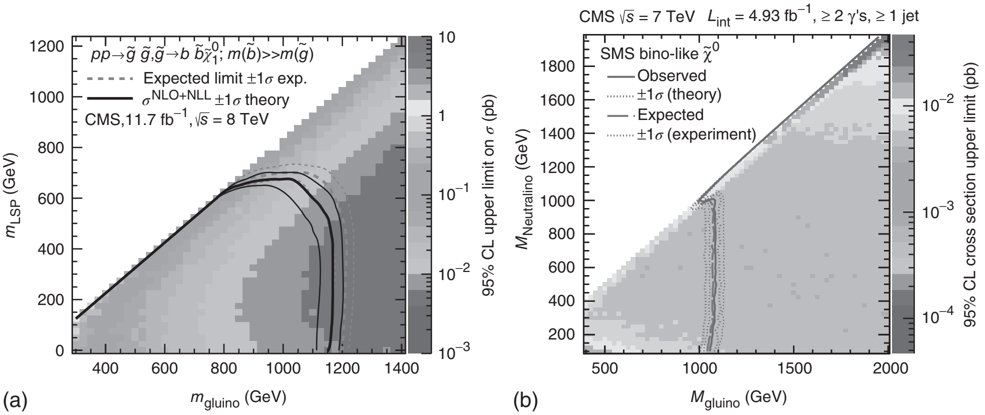


Figure 5.13 Results for simplified models in the form of 95% CL upper limits on the cross section plus overlaid exclusion contours for (a) the bottom [485] analysis as a function of the parent and LSP sparticle masses and (b) the diphoton. In (a), the blue (black) solid thick line indicates the observed exclusion assuming NLO+NLL SUSY production cross section. The blue dotted (black solid) thin lines represent the observed exclusions when varying the cross section by its theoretical uncertainty. The purple dashed thick (thin) line indicates the median ($\pm 1\sigma$) expected exclusion. (Reproduced with permission of [485, 486].) (Please find a color version of this figure on the color plates.)

Figure 5.13b shows a result of gluino-$\tilde{\chi}_1^0$ mass exclusion limit by analysing the process including two photons in the final state [486]. It was obtained by assuming direct gluino production and two photon final state as shown in Figure 5.12b. Analysis based on generalized gauge mediation model (GGM) (extension of GMSB where the lifetime is a free parameter) produced a similar result on the neutralino mass.

Analysis formalism Several new kinematic variables are implemented for analysis of LHC jet data. α_T, MT_2, razor, and so on. We explain a few to help understanding the data given by the ATLAS and CMS groups.
$\boldsymbol{\alpha_T}$: [487] In case of two-jet events,

$$\alpha_T \equiv \frac{E_{2T}}{M_{T,jj}}, \quad M_{T,jj} = \sqrt{(E_{1T} + E_{2T})^2 - (\vec{p}_{1T} + \vec{p}_{2T})^2} \tag{5.63}$$

where $E_{1T} > E_{2T}$ are the transverse energy of two jets. Background events generally tail off at $\alpha_T \simeq 0.5$ while supersymmetric events with invisible decay products can easily have larger α_T. A typical α_T plot is shown in Figure 5.14a.
MT2: [488] MT2 (or M_{T2}) analysis method shown in Figure 5.14b [489] tries to determine the mass of the primary pair-produced particles assuming that both decay into identical channels with missing neutralino (LSP). Two transverse masses are reconstructed for the pair

$$\begin{aligned} M_{T2}(m_{\tilde{\chi}}) &= \min_{\vec{p}_T^{\tilde{\chi}(1)} + \vec{p}_T^{\tilde{\chi}(2)} = \vec{p}_T^{\text{miss}}} \left[\max\left(M_T^{(1)}, M_T^{(2)} \right) \right] \\ (M_T^{(i)})^2 &= (m^{\text{vis}(i)})^2 + M_{\tilde{\chi}}^2 + 2\left(E_T^{\text{vis}(i)} E_T^{\tilde{\chi}(i)} - \vec{p}_T^{\text{vis}(i)} \cdot \vec{p}_T^{\tilde{\chi}(i)} \right) \end{aligned} \tag{5.64}$$

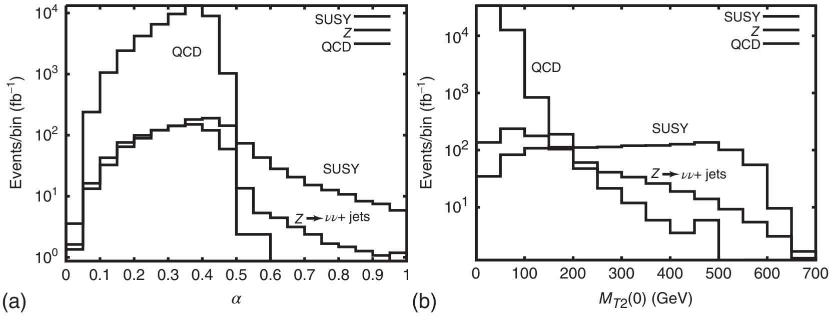


Figure 5.14 Concept of α_T and $MT2$ variables to enhance SUSY signals. The histograms are for SUSY, $Z \to \nu\nu$+jets, and QCD rates for passing the cuts, as a function of α_T (a) and $MT2$ (b). (Reproduced with permission of [487].)

where the LSP mass $m(\tilde{\chi})$ remains a free parameter. The distribution of MT2 for the correct value of $m(\tilde{\chi})$ then has an endpoint at the value of the primary particle mass. In practice, determination of MT2 may be complicated by the presence of transverse momentum arising from decays that occur upstream in the decay chain. In this case, no analytic expression for MT2 is known, but it can be computed numerically, using, for example, the results obtained in [490].

The razor analysis: The razor analysis was devised to pick up new heavy particles pair produced in the hadron collider [491, 492]. The new particles are typically produced at slightly above the threshold and hence they can be assumed at rest in the center of mass ($\tilde{q}$-COM) frame. A typical application is the pair production with subsequent two body decays of the squarks.

$$p + p \to \tilde{q}_1 + \tilde{q}_2, \quad \tilde{q}_i \to q_i + \tilde{\chi}^0 \tag{5.65}$$

The energy-momenta of the quarks $q_{i\,(i=1,2)}$ in the $\tilde{q}$-COM frame are given by

$$p_q = -p_{\tilde{\chi}}, \quad E_q = |p_q| = \frac{M_\Delta}{2} \equiv \frac{M_{\tilde{q}}^2 - M_{\tilde{\chi}}^2}{2M_{\tilde{q}}} \tag{5.66}$$

where $M_{\tilde{q}}$ and $M_{\tilde{\chi}}$ are the squark and LSP masses, respectively. One naturally wants to reconstruct the energy-momenta of the quarks in $\tilde{q}$-COM frame. However, to obtain a correct Lorentz parameter β_{CM} to boost from the laboratory frame to the $\tilde{q}$-COM system is not possible because of undetected particles. The strategy of the razor analysis is to approximate this unknown frame with a razor frame, which need not be specified in the following arguments[8]

---

[8] Actually, the razor frame is defined by a Lorentz boost factor $\beta_R = (p_{1z} + p_{2z})/(E_1 + E_2)$. The energy of the jets in the R-frame is expressed as [491]

$$E_{1,2}^R = \gamma_R\left(E_{1,2} - \beta_R p_{1z,2z}\right) = \frac{1}{2}\left[M_R \pm (\vec{p}_{1T} - \vec{p}_{2T}) \cdot (\vec{E}_T^{miss})\right] \tag{5.67}$$

They do not exactly coincide with $M_R/2$ but $E_1^R + E_2^R = M_R$.

Two mass variables and one parameter that can be reconstructed from observables are defined.

$$\begin{aligned} M_R &= \sqrt{(E_1 + E_2)^2 - (p_{1z} + p_{2z})^2} \\ M_T^R &= \sqrt{\frac{p_T^{\text{miss}}(p_{1T}+p_{2T}) - \vec{p}_T^{\text{miss}} \cdot (\vec{p}_{1T}+\vec{p}_T^{\text{miss}})}{2}} \\ R &\equiv \frac{M_T^R}{M_R} \end{aligned} \tag{5.68}$$

where $E_i, p_{iz}(i = 1, 2)$ are energy and longitudinal momentum along the beam axis of the ith jet that results from hadronization of the squarks and suffix T and "miss" denote transverse and missing momentum. The longitudinal momenta are determined from the massless on-shell conditions for each particle in the jet. The parameter R is referred to as the *razor parameter*. M_T^R has kinematical endpoint at M_R (the Jacobian peak), thus the value of R has a maximum value of approximately one. Notice that both variables are invariant under Lorentz transformation along the z (i.e., beam) axis and in the $\tilde{q}$-COM frame $M_R = M_\Delta$ for squarks at rest. On event-by-event basis, M_R distributes around M_Δ because of fluctuations in undetected energies. Therefore, it is an indicator for $\gamma_{\text{CM}} M_\Delta$.

For signal events, M_R peaks around M_Δ and a distribution in R peaks around ~ 0.5, in sharp contrast with QCD multijet background events that are exponentially suppressed. Figure 5.15a shows how a hypothetical sparticle signal appear in M_R distribution and Figure 5.15b shows exponential background suppression by R-cut. Therefore, by performing 2D analysis in $M_R - R$ plane, one can extract the signal efficiently with minimum backgrounds.

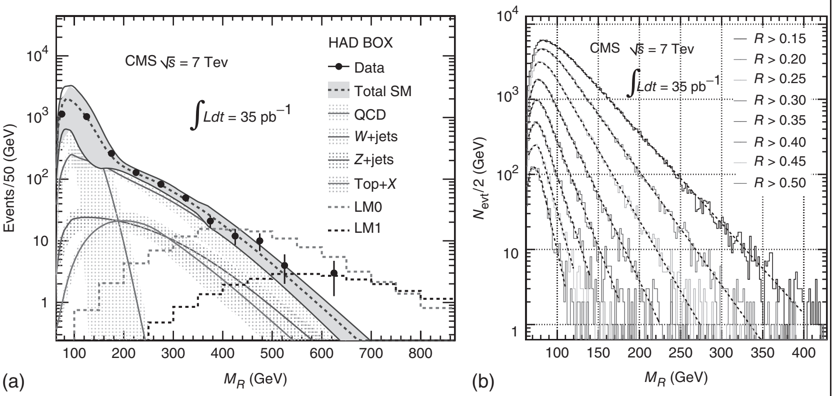


Figure 5.15 (a) The M_R distributions with $R > 0.5$ for data (points) and backgrounds (curves) on logarithmic scales. The bands show the uncertainties of the background predictions. The corresponding distributions for SUSY benchmark models LM1 [493] with $M_\Delta = 597$ GeV and LM0 [494] with $M_\Delta = 400$ GeV are overlaid. (b) M_R distributions for different values of the R threshold for data events in the QCD multijet control sample. Fits of the M_R distribution to an exponential function and an asymmetric Gaussian at low M_R are shown as dotted black curves. (Reproduced with permission of [495].)

For analyses using mSUGRA/CMSSM, five parameters have to be specified. If no signals for new physics were detected, analysis result of the experimental data gives excluded regions on $x - y$ plane where x and y are two of the five parameters with others fixed at some typical values. The current excluded region on $(m_{\tilde{g}}, m_{\tilde{q}})$ and $(m_{1/2}, m_0)$ obtained by the LHC at 7 and 8 TeV energy is given in Figure 5.16 [8, 496, 498]. Here, $\tan\beta = 30$, $A_0 = -2m_0$, $\mu > 0$ are used. The gluino masses below 1100 GeV are excluded for all squark masses. If the squark mass and that of the gluino are assumed equal, the limit is about 2000 GeV. This is a drastic improvement compared to 380 GeV obtained at Tevatron [499, 500].

5.7.6
Stop

In many scenarios, the stop is the lightest squark. Moreover, the stop production exhibits characteristic decay modes not seen for squarks of the first and second generations. Because of this, dedicated searches for the stops can be performed. The search, for instance, can look into direct production of stop pairs followed by decays into a top and the lightest supersymmetric particle $\tilde{\chi}_1^0$.

$$\begin{aligned}
&pp \to \tilde{t}_1\tilde{t}_1 \\
&\tilde{t}_1 \to \begin{cases} t + \tilde{\chi}_1^0 \to bW\tilde{\chi}_1^0 & \text{if } m_t < m_{\tilde{t}_1} \\ b + \tilde{\chi}_1^\pm \to bW\tilde{\chi}_1^0 & \text{if } m_t > m_{\tilde{t}_1} \end{cases} \\
&W \to \begin{cases} q_i\bar{q}_j & \text{all hadronic} \\ l\nu_l & \text{leptonic} \end{cases}
\end{aligned} \tag{5.69}$$

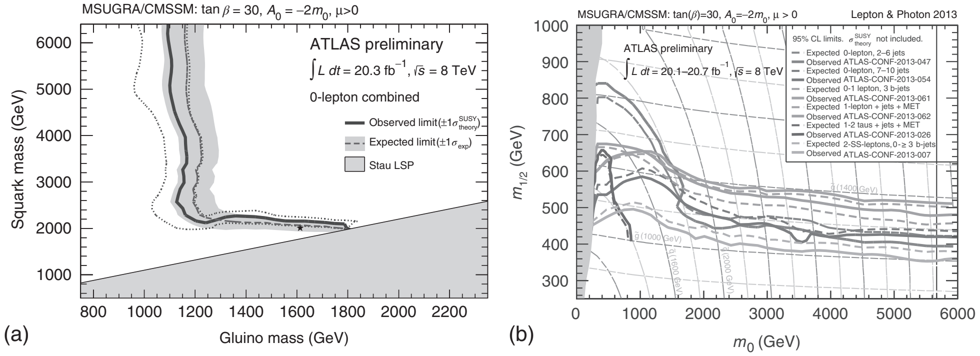


Figure 5.16 Exclusion limits for MSUGRA/CMSSM models with $\tan\beta = 30$, $A_0 = -2m_0$, and $\mu > 0$ presented (a) in the $m_{\tilde{g}} - m_{\tilde{q}}$ plane and (b) in the $m_0 - m_{1/2}$ plane. In (a), the blue dashed lines show the expected limits at 95% CL, with the light (yellow) bands indicating the 1σ excursions due to experimental and background-theory uncertainties. In (b), the observed limits are indicated by medium (maroon) curves, where the solid contour represents the nominal limit, and the dotted lines are obtained by varying the signal cross section by the theoretical scale and PDF uncertainties. (Reproduced with permission of [496, 497].) (Please find a color version of this figure on the color plates.)

where $\tilde{t}_1$ is the lightest $\tilde{t}$ mass eigenstates. The topology of the above decay channels is 6 hadronic jets plus large missing transverse energy for W decaying all hadronically or $1 \sim 2$ leptons + b jets +-missing energy. Figure 5.17 shows a collection of analyses carried out by the ATLAS group [501, 502]. A region in the $\tilde{t}$ versus $\tilde{\chi}_1^0$ mass plane bounded by $m_{\tilde{t}_1} \lesssim 660$ GeV for $m_{\tilde{\chi}_1^0} \lesssim 250$ GeV is excluded.

5.7.7
R-hadrons

R-hadrons are hypothetical hadrons composed of a supersymmetric particle and at least one quark. Generally, supersymmetric particles are so separated in mass that their decays are very fast with an exception of LSP. But in certain circumstances the *R*-hadrons may be formed by the lightest colored sparticle. A prime example is the gluino in the split SUSY [40–42]. Its mass scale is typically approximately teraelectronvolt while all other masses of sfermions or new bosons are at a much higher scale. The *R*-parity conservation forbids the gluino to decay directly to quarks and/or gluons. It has to go through other colored squarks that are much heavier. Depending on the mass, the gluino can have much longer lifetime compared to the normal. If its mean lifetime is longer than the typical hadronization time scale, the QCD bound states can be formed with SM particles (quarks and gluons) in analogy

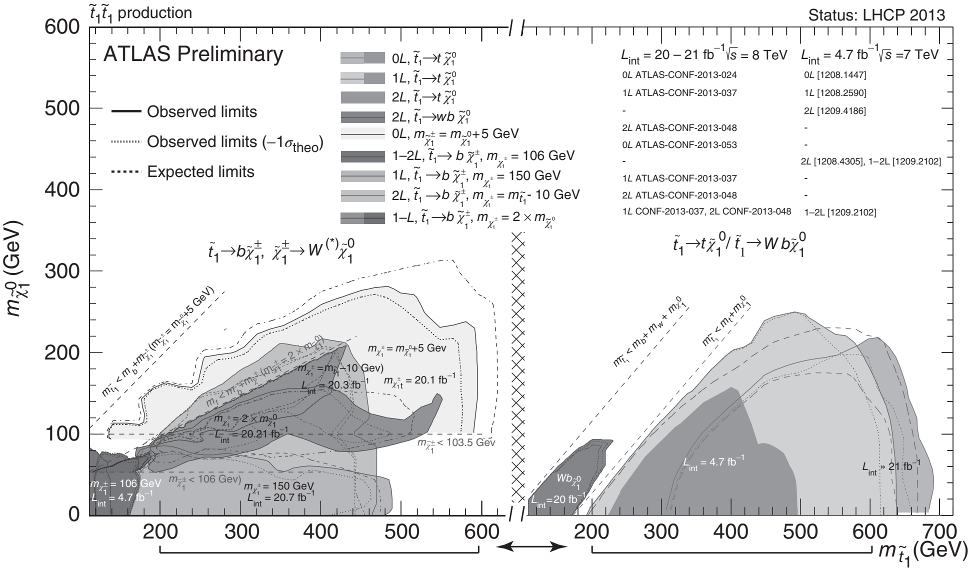


Figure 5.17 Searches for the pair-produced stop with both stops subsequently decaying to top and $\tilde{\chi}_1^0$ excluded the regions in red color. Also shown in blue color are the regions excluded by looking into 1 or 2 leptons + b-jets. Reach for $b\tilde{\chi}_1^+$ analysis is very dependent on $\tilde{\chi}_1^+$ mass but some sensitivity to regions not probed by $t\tilde{\chi}_1^0$ (avoid top in decay chain). (Reproduced with permission of [501, 502].) (Please find a color version of this figure on the color plates.)

with the ordinary hadrons. Another example is provided by the stop, which is the lightest colored particle in most SUSY models.

The *R*-hadron may have a lifetime long enough to be detected by time-of-flight measurement or its decay vertex detached from the primary interaction point so as to be detectable by microvertex detectors. If the *R*-hadron is charged, it can ionize the detector material. Different ionization pattern other than the minimum ionization loss would tell the mass and charge of the traversing particle. Yet another characteristic pattern that only the *R*-hadron will have is the possibility of charge exchange reaction with ordinary matter. This can happen because the *R*-hadron is a composite and exchange of its constituent quark, say an up-quark with a down-quark in the detector material, will result in charge variation that can be observed by looking at the ionization loss pattern. Results on *R*-hadron searches are generally given as another constraints on lower limits on the mass of gluinos, squarks, or stops. No signals were detected so far [477, 503]. Figure 5.18 shows the upper limit of the cross sections obtained by the CMS group, which are compared with various models.

5.7.8
Gravitino

In the GMSB model, the gravitino $\tilde{G}$ is the LSP. The cross section for associate production of gravitino–squark ($pp \to \tilde{G}\tilde{q} + X$) and gravitino–gluino ($pp \to \tilde{G}\tilde{g} + X$) processes are given in [507]. The cross section depends on $m_{\tilde{G}}$ as $\sigma \sim 1/m^2_{\tilde{G}}$, which provides means to determine lower bounds on $m_{\tilde{G}}$. The decay of the gluino or squark into a gravitino and a gluon ($\tilde{g} \to \tilde{G} g$) or a gravitino and a quark ($\tilde{q} \to \tilde{G} q$), respectively, dominates and the final state is characterized by

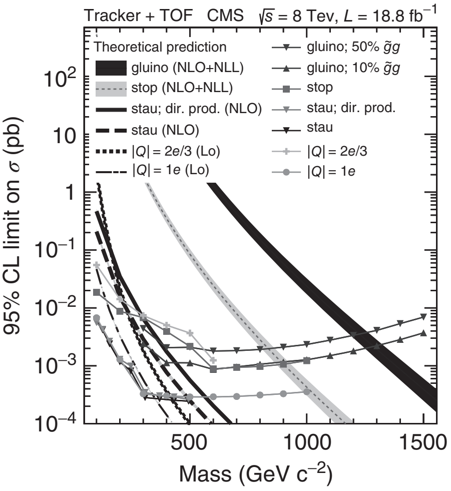


Figure 5.18 Upper cross section limits at 95% CL on various signal models for the tracker + TOF analysis. (Reproduced with permission of [504].)

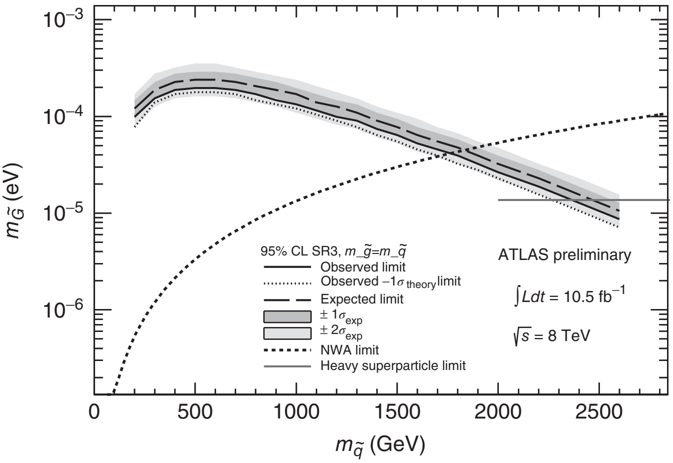


Figure 5.19 Limits on the gravitino mass: observed (solid line) and expected (dashed line) 95% CL lower limits on the gravitino mass as a function of the squark mass for degenerate squark/gluino masses. The shaded bands around the expected limit indicate the expected $\pm 1\sigma$ and $\pm 2\sigma$ ranges of limits in the absence of a signal. The short dashed line going upwards defines the validity of the narrow-width approximation used in the analysis. The solid short line at the right end denotes the current limit from LEP [505] on the gravitino mass assuming very heavy squarks/gluino. (Reproduced with permission of [506].)

the presence of a pair of gravitino that escapes detection leading to an energetic jet and missing energy. Figure 5.19 shows a result obtained by [506] by looking for a mono jet + missing energy topology.

Gravitino masses below 1×10^{-4} eV (4×10^{-5} eV) are excluded at 95% CL for squark/gluino masses of 500 GeV (1.7 TeV).

5.8
Current Status

A summary of the SUSY particle searches in direct productions as of summer in 2013 obtained by the ATLAS group is given in Figure 5.20. The CMS group produced a similar result [390].

Other constraints on the supersymmetric models can be placed using the small value of the Higgs mass as discussed in Chapter 1.

A lot of evidence indicates that the low-energy SUSY may not be minimal after all. The benchmark spectrum given in Figure 5.6 is already outmoded. Theorists have began thinking seriously high-scale or nonminimal scenario of SUSY. The notion of naturalness is also in doubt. However, at the time of writing this book (summer, 2013), the LHC was still on its way to achieve the designed luminosity $5 \times 10^{34}\ \text{cm}^{-2}$. The author would like to close this chapter by mentioning that there still remains a possibility of the "natural SUSY scenario" [509–511].

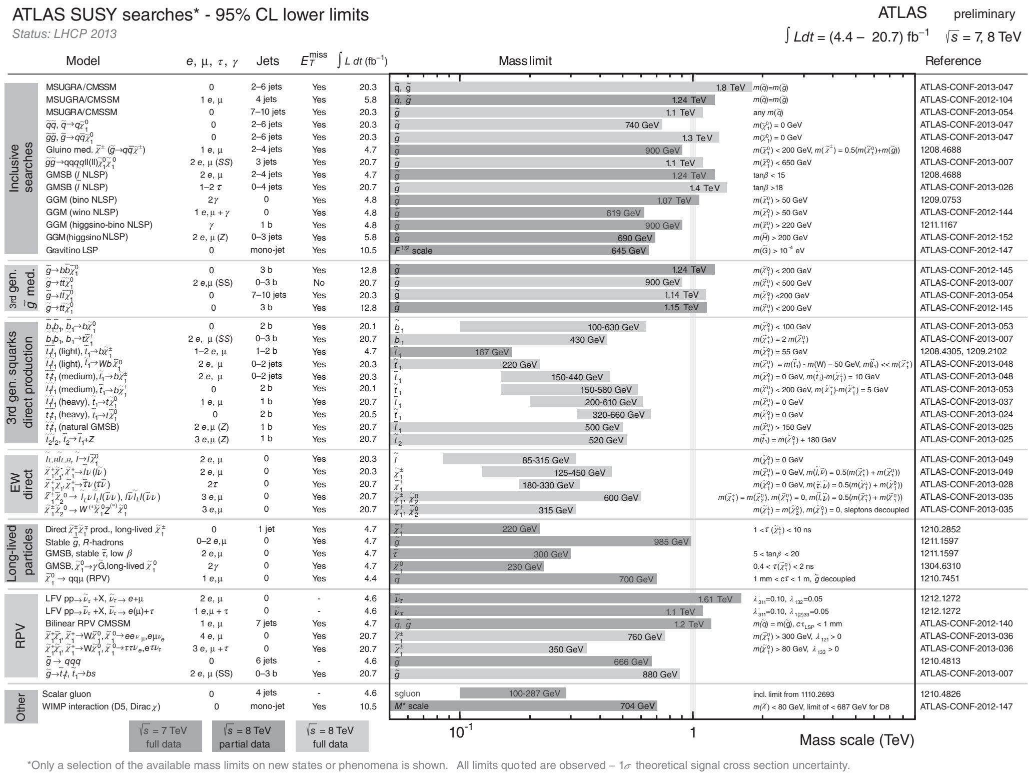


Figure 5.20 ATLAS susy mass reaches. (Reproduced with permission of [508].)

The essence of the natural SUSY can be summarized by the following relation. [see Eq. (5.13)]

$$\frac{m_Z^2}{2} = \frac{m_{H_d}^2 - m_{H_u}^2 \tan^2\beta}{\tan^2\beta - 1} - \mu^2 \overset{\tan^2\beta\gg1}{\simeq} -\mu^2 - m_{H_u}^2 \tag{5.70}$$

The requirement for the naturalness means no fine-tuning. In other words, each term on the right-hand side should be of the EW scale, that is, no large compensation between the two. It immediately follows that the higgsinos should not be too heavy because their mass is controlled by μ. The stops ($Q_3 \sim \tilde{t}_L$, $u_3 \sim \tilde{t}_R$) also cannot be too heavy because the radiative corrections to $m_{H_u}^2$ are given by [37]

$$\delta m_{H_u}^2|_{\text{stop}} = -\frac{3}{8\pi^2} y_t^2 (m_{Q_3}^2 + m_{u_3}^2 + |A_t|^2) \ln\left(\frac{\Lambda}{\text{TeV}}\right) \tag{5.71}$$

where Λ denotes the scale at which SB are mediated to the supersymmetric SM. Equation (5.71) together with naturalness requirement severely affects the stop mass from above (several hundred gigaelectronvolts to 1~2 TeV depending on the parameter setting). The gluino can also be affected. It should be above the stop but less than a few teraelectronvolt because it contributes to the stop mass at one-loop level, which, in turn, contributes to the Higgs mass at two-loop level. The level structure of the natural SUSY is depicted in Figure 5.21.

The masses of the rest of the superpartners, including the squarks of the first two generations, are not important for naturalness and can be much heavier than the present LHC reach (See also discussions on split SUSY in Section 1.4.3.).

The detection of the higgsinos is hard at LHC but the stops could be searched for at LHC as was already demonstrated in Figure 5.17.

We expect more data from LHC and later from future linear colliders. They will help focusing the direction of the new SUSY models. Finally, we show a region on

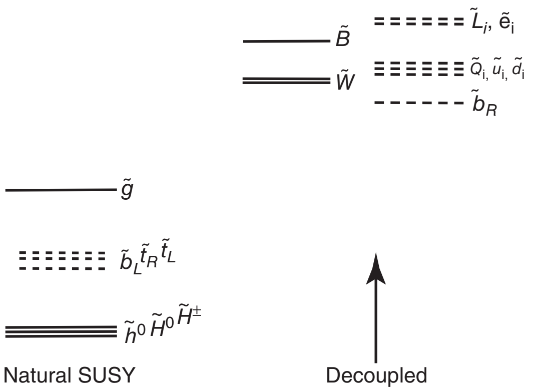

$\tilde{B}$
$\tilde{L}_i$, $\tilde{e}_i$
$\tilde{W}$
$\tilde{Q}_i, \tilde{u}_i, \tilde{d}_i$
$\tilde{b}_R$
$\tilde{g}$
$\tilde{b}_L \tilde{t}_R \tilde{t}_L$
$\tilde{h}^0 \tilde{H}^0 \tilde{H}^\pm$
Natural SUSY
Decoupled

Figure 5.21 Natural electroweak symmetry breaking affects the superpartners on the left to be light. Meanwhile, the superpartners on the right can be heavy, $M \gg 1$ TeV, without spoiling the naturalness. [509, 510, 512]

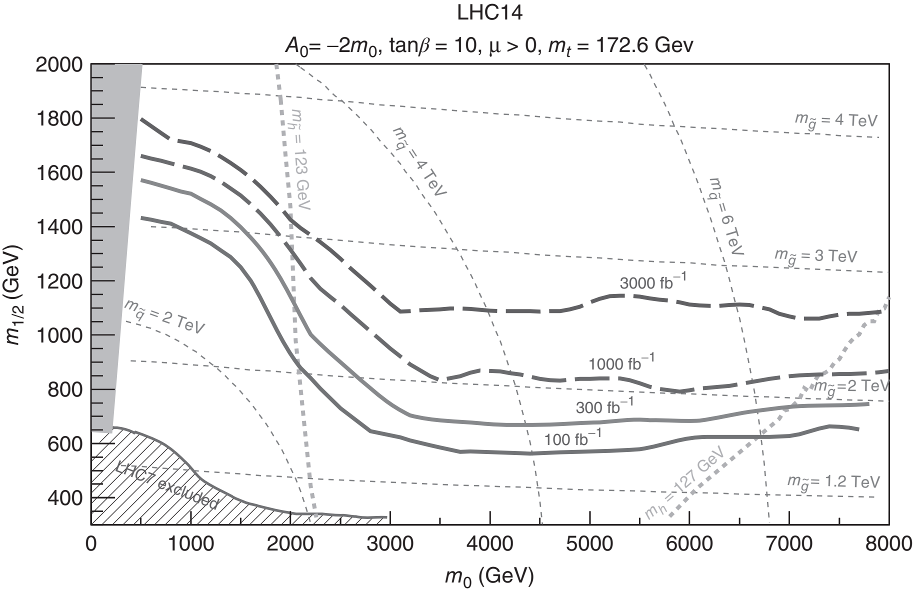


Figure 5.22 Expected SUSY reach for LHC at 14 TeV for integrated luminosities of 100, 300, 1000, and 3000 fb^{-1}. The signal is observable if it falls below the curve for the corresponding integrated luminosity. The fixed mSUGRA parameters are $A_0 = -2m_0$, $\tan\beta = 10$, and $\mu > 0$. Gluino and squark mass contours are shown by the dashed, dark gray curves. We also show contours of m_h = 123 and 127 GeV. The shaded gray area on the left side of the figure is excluded because the stau becomes the LSP. The shaded region on the lower left side is excluded by SUSY searches at LHC at 7 TeV. (Reproduced with permission of [511].)

$m_0 - m_{1/2}$ plane that can be searched for when the full luminosity is realized at 14 TeV [Figure 5.22].

6
Extra Dimension

6.1
Introduction

One of the outstanding issues in particle physics is the hierarchy problem. There are two kinds: one is why gravity is so weak compared to other fundamental forces and the other is why the top quark is so heavy relative to others. The former is referred to as the *gauge* or *big hierarchy* and the latter as the *flavor* or *mass hierarchy*.[1] One outcome of the big hierarchy is the fine-tuning problem discussed in the Higgs and grand unified theory chapters. Several ideas to solve it were proposed; among them, the supersymmetry was the most thoroughly investigated. However, the fine-tuning problem is a technical matter that is associated with the hierarchy. The supersymmetry does not address why such a hierarchy exists at all. It does not provide solutions to the mass hierarchy either. The extra dimension (ED) may possibly provide an answer to both questions.

The idea of the ED is old [513]. Immediately after Einstein's general relativity was published, Kaluza and Klein proposed a model to unify gravity and electromagnetism by extending the general relativity to five-dimensional spacetime [514]. It was forgotten for a long time but was revived with the advent of the string theory. The basic notion of the string theory is very similar to that of the Kaluza–Klein (KK) theory in that it tries to unify all forces in the framework of general relativity in 10- or 11-dimensional spacetime. The difference is that the string theory treats strings rather than particles in the framework of quantized local supersymmetry.

Gravity is different from other forces in many respects. The graviton is a quantum excitation of the spacetime metric tensor and hence has spin 2. The vacuum expectation value (VEV) of the graviton field is the classical spacetime whose metric defines the background vacuum. Its coupling strength is extremely weak. If we want to treat gravity on equal footing with other forces, we have to go to the Planck energy ($M_{\text{Pl}} = 1/\sqrt{G_N} \sim 1.22 \times 10^{19}$ GeV), which is some 10^{17} orders apart in scale from the electroweak scale ($(G_F)^{-1/2} \sim 250$ GeV). This is the (big) hierarchy problem.

---

[1] We also have little hierarchy problem that we discussed in Section 1.4.

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

The ED offers a possibility for solving the hierarchy problem [515]. Classically, gravity has its roots in the geometry of spacetime. If one wants to modify it, it is natural to extend its dimension. Suppose that the spacetime dimension is $D = 4 + \delta\,(\delta = 1, 2, \cdots)$, gravity spreads in $D - 1$ dimensional space. Then, Gauss's law tells that the gravitational force flux (and hence its strength) decreases $f \sim 1/r^{2+\delta}$.

$$f(r) = \begin{cases} -G_N \dfrac{m_1 m_2}{r^2} \equiv -\dfrac{1}{M_{\text{Pl}}^2} \dfrac{m_1 m_2}{r^2} & D = 4 \text{ dimensional} \\ -G_D \dfrac{m_1 m_2}{r^{2+\delta}} \equiv -\dfrac{1}{M_D^{(2+\delta)}} \dfrac{m_1 m_2}{r^{(2+\delta)}} & D = 4 + \delta \text{ dimensional} \end{cases} \tag{6.1}$$

G_D is the fundamental gravitational constant in D-dimensional space corresponding to $G_N \equiv 1/M_{\text{Pl}}^2$ in 3D space. Notice that the gravitational constant in higher dimension has different mass dimension.

We want to investigate a possibility that the energy scale of the fundamental constant M_D may not be too different from the electroweak scale that we conveniently set at 1 TeV. The gravity under consideration should not contradict Newton's inverse square law. This can be realized if the gravitational force in the ED saturates at some value $R \sim 1/M_D$. Then, for $r \gg R$, the power of the force propagation decreases to $\sim 1/r^2$, recovering the usual Newton's law.

The dimensional argument constrains the force as

$$f = G_D \frac{m_1 m_2}{r^{2+\delta}} \xrightarrow{r \gg R} \frac{1}{M_D^2} \frac{m_1 m_2}{r^2} \frac{1}{(M_D R)^\delta} {}^{2)} = G_N \frac{m_1 m_2}{r^2} = \frac{1}{M_{\text{Pl}}^2} \frac{m_1 m_2}{r^2} \tag{6.2a}$$

$$\therefore \quad R = \left(\frac{M_{\text{Pl}}^2}{M_D^{2+\delta}} \right)^{\frac{1}{\delta}} \simeq \left(\frac{M_D}{1\,\text{TeV}} \right)^{\frac{-(\delta+2)}{\delta}} \times (10)^{\left(\frac{32}{\delta}\right) - 17} \text{cm}$$

$$= \begin{cases} \sim 0.4 \text{ light years} & \delta = 1 \\ \sim 1 \text{ mm} & \delta = 2 \\ \sim 5 \times 10^{-7} \text{cm} & \delta = 3 \\ \sim 10^{-9} \text{cm} & \delta = 4 \end{cases} \tag{6.2b}$$

where m_1, m_2 are masses of the particles introduced conveniently, which disappear from the final expression. For $r \ll R$, $f \sim r^{-(2+\delta)}$, the force violates the Newton's inverse square law. From Eq. (6.2b), we see that $\delta = 1$ and possibly $\delta = 2$ are ruled out. However, Newton's law was never tested at or below the scale of 0.1 mm. Hence, $\delta \geq 3$ is allowed (see Figure 6.1). In other words, the fundamental gravitational energy scale to be ~ 1 TeV is not excluded if one allows EDs. Here lies a possibility that the hierarchy after all may not exist. Notice that other particles in the Standard Model (SM) are not allowed to reside in such a large dimensional space, because the SM was tested at the energy scale up to 1 TeV or $\sim 10^{-18}$ cm. If one allows the SM particles to reside in the ED, its size has to be smaller than $\sim \text{TeV}^{-1}$.

---

2) The exact definition of M_D that we will use later in discussing the gravity is given by $M_D^{2+\delta} R^\delta = 1/8\pi G_N$, where the ED is compactified on a circle of radius R.

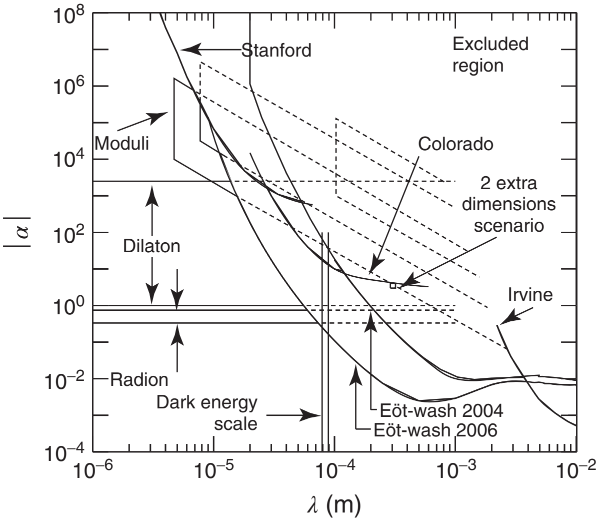


Figure 6.1 Experimental limits on α and λ of the gravitational potential, which is parameterized as $V(r) = (-G_{\text{Pl}} m_1 m_2 / r)(1 + \alpha e^{-r/\lambda})$. Large values of α, λ violate Newton's law. The upper right region is experimentally excluded. (Reproduced with permission of [7, 516–518].)

The idea of an ED has triggered many models and their experimental tests have been proposed. In some models like the one we described above, only the graviton resides in ED as well as in the Minkowski space and the ordinary particles are confined in the four-dimensional Minkowski space. The whole spacetime is referred to as the *bulk*. In other models, all the particles may reside in the bulk (referred to as the *universal extra-dimension model*) as well, or may reside in a limited subspace inside the bulk (referred to as the *brane*).

Treating particles other than the graviton differently may seem odd at first glance. The idea was inspired by the D-brane of the string theory where endpoints of open strings reside [519]. In the case of a one-dimensional string that we learn in classical mechanics, two kinds of boundary conditions are known: the Dirichlet and Neumann conditions.[3] In the string theory, the endpoint with the Dirichlet condition can be a p-dimensional space. If $p = 1$, the endpoint is a point in one dimension. For $p \geq 2$, it is a point in p-dimensional plane (or brane), which is generically called *D-brane* (or Dp-brane). In QCD, we know that the quarks are confined because the color force behaves like a string and the quarks can be regarded as endpoints of the strings. In the string theory, ordinary particles are considered as endpoints of open strings, whereas the graviton is a closed string

---

[3] Denoting the amplitude of the string oscillation as $A(x)$, $a \leq x \leq b$, two kinds of boundary conditions can be applied.

$$\begin{array}{ll} A(a) = A(b) = 0 & \text{Dirichlet condition} \\ \left.\frac{\partial}{\partial x} A(x)\right|_{x=a} = \left.\frac{\partial}{\partial x} A(x)\right|_{x=b} = 0 & \text{Neumann condition} \end{array} \qquad (6.3)$$

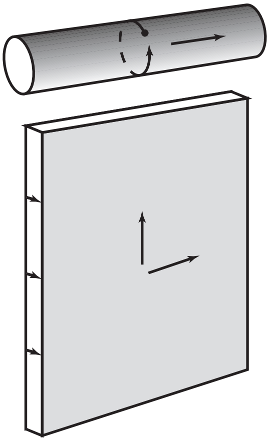


Figure 6.2 A small dimension of space may have two different forms. It may be curled up to form a cylinder, and the dimension transverse to its axis forms a closed cycle. But it may also form an interval line, the thickness of a box. It is impossible to go out of the box, as there is nothing there, not even vacuum. It is really impossible to visualize a real extra dimension within our three-dimensional space. What is usually done is to omit one dimension of our space and to depict it in the extra-dimensional space. The direction in our space (long arrows in the drawing) is said to be longitudinal, whereas the direction in the extra dimension (circle or short arrows in the drawing) is said to be transverse to our space.

that can fly freely in the bulk. Then, ordinary particles are fixed at the p-brane but move freely inside it, thus justifying the notion that the graviton and other particles may reside in different spaces. Note that it is customary not to count the time dimension for the brane; hence, our four-dimensional spacetime is referred to as a *three-brane*. This is a bit confusing but is conventional.

There have been no experimental or observational evidences that the ED exists, so if does exist it is not the same as the known 3D space. It may be limited and finite like those depicted in Figure 6.2.

The original proposal by Kaluza and Klein and the traditional notion of the string theory were to consider the size to be on the order of the Planck size defined by $\hbar/M_{\mathrm{Pl}}c \equiv \sqrt{\hbar G_N/c^3} \sim 1.6 \times 10^{-36}$ m or, equivalently, $M_{\mathrm{Pl}}^{-1} \sim (10^{19}\ \mathrm{GeV})^{-1}$. The notion that the size of ED may be as large as TeV^{-1} was contemplated by [520]. But it was [515] who considered large extra dimensions (LEDs) as a solution to the hierarchy problem that made the theme a popular topic in phenomenology.4) Warped ED [521, 522], universal ED [523], and AdS/CFT correspondence [524] provided other new possibilities and made the notion of ED very attractive.

The field of ED is vast and developing fast. While it has many desirable features, viable models in a self-consistent mathematical framework are far from realization. In the following, we limit ourselves to discussions on some of their salient features and possible experimental signals that we may be able to obtain in the near future.

6.2
KK Tower

We first discuss how a higher dimensional phenomenon looks like in our 4D world. For the sake of simplicity, we limit ourselves to one ED in the following

---

4) Referred to as the *Arkani-Hamed–Dimopoulos–Dvali* (ADD) LED or simply ADD. In this model, only the graviton resides in ED, which is flat and compactified.

discussions. The momentum of a particle in ED appears in the disguise of a mass in our space

$$E = \sqrt{\boldsymbol{p}^2 + p_5^2 + m^2} \equiv \sqrt{\boldsymbol{p}^2 + m_{\mathrm{KK}}^2} \tag{6.4}$$

where p_5 is the momentum of the particle in the fifth dimension. p_5 is continuous if the space extends to infinity. In our case, however, the ED is compactified on a circle with radius R, and hence the spectrum is discrete but with an infinite number of levels. Therefore, we have a whole series of particles having different masses, which is referred to as the *Kaluza–Klein* (abbreviated as KK) tower.

Let us quantify our arguments. We shall denote the five-dimensional (hereafter denoted as 5D) spacetime coordinates as $X^M = (x^\mu, x^5)$ $(M = 0 \sim 3, 5)$ and assume that $x^5 \equiv y$ is confined on a circle $0 \leq y < 2\pi R$ and any function satisfies $\phi(x, y + 2\pi R) = \phi(x, y)$.[5] The metric for the background vacuum is defined as

$$\begin{aligned} ds^2 &= \eta_{\mathrm{MN}} dX^M dX^N = \eta_{\mu\nu} dx^\mu dx^\nu - dy^2 \\ \eta_{\mathrm{MN}} &= \mathrm{diag}(1, -1, -1, -1, -1) \end{aligned} \tag{6.5}$$

The action of a free (massless) scalar field Φ in 5D is

$$\begin{aligned} S^{(5\mathrm{D})} &= \int d^5X \left(\frac{1}{2} \partial_M \Phi(X) \partial^M \Phi(X) \right) \\ &= \int d^4x \int_0^{2\pi R} dy \frac{1}{2} \left[\partial_\mu \Phi(x, y) \partial^\mu \Phi(x, y) - \partial_y \Phi(x, y) \partial_y \Phi(x, y) \right] \end{aligned} \tag{6.6}$$

where we used abbreviation X to denote X^M and x to denote x^μ. Note that, as the mass dimension of the action and spacetime coordinates are 0 and -1, respectively, the scalar field Φ has mass dimension 3/2 in 5D. As the ED is compactified, the field value is periodic in y coordinates, $\Phi(x, y + 2\pi R) = \Phi(x, y)$. Therefore, we can expand it in Fourier series as

$$\begin{aligned} \Phi(x, y) &= \frac{1}{\sqrt{2\pi R}} \sum_{n=-\infty}^{+\infty} \varphi^{(n)}(x) e^{\frac{iny}{R}} \\ &= \frac{1}{\sqrt{2\pi R}} \phi^{(0)}(x) + \frac{1}{\sqrt{\pi R}} \sum_{n=1}^{\infty} \left\{ \phi^{(+n)}(x) \cos\left(\frac{ny}{R}\right) + \phi^{(-n)}(x) \sin\left(\frac{ny}{R}\right) \right\} \end{aligned} \tag{6.7}$$

Reality of Φ implies $\varphi^{(-n)} = \varphi^{(n)\dagger}$, and we have introduced real scalar fields $\phi^{(\pm n)}$ where

$$\varphi^{(0)} = \phi^{(0)}, \qquad \varphi^{(\pm n)}\,(n > 0) = \frac{1}{\sqrt{2}} \left(\phi^{(+n)} \mp i\phi^{(-n)} \right) \tag{6.8}$$

---

[5] In mathematical language, we say a set of real numbers $y \in \mathscr{R}$ with the constraint is equivalent to that on a circle which is denoted as $y \in S^1$.

Substituting Eq. (6.7) into Eq. (6.6) and integrating over y, we obtain

$$\begin{aligned}S^{(4\mathrm{D})} &= \int d^4x \frac{1}{2}\left[\sum_n \left\{\partial_\mu \varphi^{(-n)}\partial^\mu \varphi^{(n)} - \frac{n^2}{R^2}\varphi^{(-n)}\varphi^{(n)}\right\}\right] \\ &= \int d^4x \frac{1}{2}\left[\partial_\mu \phi^{(0)}\partial^\mu \phi^{(0)} + \sum_{n=1}^{\infty}\left\{\partial_\mu \phi^{(\pm n)}\partial^\mu \phi^{(\pm n)} - \frac{n^2}{R^2}\phi^{(\pm n)}\phi^{(\pm n)}\right\}\right]\end{aligned} \tag{6.9}$$

The first term in the second line describes the conventional massless scalar field. But the second term is an addition we obtained by going to 5D. From the 4D point of view, we have acquired an infinite series (KK tower) of particles with masses $m_n = n/R$. That is, the field theory in the compactified extra-dimensional space is equivalent to that in 4D containing an additional infinite series of massive KK towers. This $S^{(4\mathrm{D})}$ is referred to as the *4D effective action*.

If the field Φ had the intrinsic mass m_0 in 4D, then its KK tower in 5D, 6D,… will have mass

$$m^2 = m_0^2 + \frac{n_5^2}{R_5^2} + \frac{n_6^2}{R_6^2} + \cdots \tag{6.10}$$

where $n_5, n_6, \ldots$ are the indices to denote excited levels, and R_5, $R_6, \ldots$ denote the size of the fifth, sixth,… EDs. If the size of R is small or, equivalently, the mass n/R is large such that it goes beyond the presently accessible scale, we will not be able to observe the KK tower. Conversely, if the size R is reasonably large and within the reach of the accelerator energy, we should be able to observe the KK tower and hence the sign of the ED.

6.2.1
Effective Coupling Strength in 4D

What will happen to the strength of the coupling constant? Consider a point source with coupling strength g_5. The field equation with a static point source can be written as

$$\partial_M \partial^M \Phi(x^\mu, y) = g_5 \delta^3(x)\delta(y)\,^{6)} \tag{6.11}$$

Note that, as the mass dimension of Φ is 3/2, the coupling constant g_5 has mass dimension $-1/2$. Substituting Eq. (6.7) in Eq. (6.11) and integrating over y, we obtain

$$\sqrt{2\pi R}\,\partial_\mu \partial^\mu \varphi^{(0)}(x^\mu) = g_5 \delta^3(x) \tag{6.12}$$

---

6) Consider a scalar field φ in 3D, which interacts with another field ψ with the coupling strength g, that is, the interaction Lagrangian has the form $g\overline{\psi}\psi\varphi$ (imagine, e.g., φ = pion, ψ = nucleon). If the nucleon is localized at $x = 0$, it acts as a point source and the field satisfies the equation $(-\nabla^2 + m^2)\varphi = g\delta^3(x)$, which is a 3D static version of the Klein–Gordon equation in Eq. (6.12).

The static solution will be given by

$$\varphi^{(0)} = \frac{1}{\sqrt{2\pi R}} \times \frac{g_5}{4\pi}\frac{1}{r} \equiv \frac{g_4}{4\pi}\frac{1}{r} \qquad \to g_4^2 = \frac{g_5^2}{2\pi R} \tag{6.13}$$

This means that the effective coupling strength in 4D is reduced by the volume factor $V = 2\pi R$ of the ED.

Another way of looking the coupling strength in 4D is to work in 4D from the beginning using the effective 4D Lagrangian of Eq. (6.9). The equation of fields yields

$$\begin{aligned} \partial_\mu \partial^\mu \phi^{(0)} &= 0, \\ (\partial_\mu \partial^\mu + m_n^2)\phi^{(\pm n)} &= 0, \quad m_n = \frac{n}{R} \end{aligned} \tag{6.14}$$

As is well known, the static solution of the massive Klein–Gordon equation is a Yukawa potential. Therefore, considering the potential created by a point source of strength g_4, the effect of additional KK towers can be calculated as

$$\begin{aligned} V_{\text{eff}} &= \frac{g_4^2}{4\pi r}\left[1 + 2\sum_{n=1}^{\infty} e^{-\frac{nr}{R}}\right] = \frac{g_4^2}{4\pi r}\left[1 + 2\frac{e^{-\frac{r}{R}}}{1 - e^{-\frac{r}{R}}}\right] \\ &\to \begin{cases} \dfrac{g_4^2}{4\pi r} & r \gg R \\ 2\pi R \times \dfrac{g_4^2}{4\pi^2 r^2} & r \ll R \quad ^{7)} \end{cases} \end{aligned} \tag{6.15}$$

The factor 2 in the summation accounts for n in Eq. (6.7) running from $-\infty$ to $+\infty$. The potential for the case $r \gg R$ is the standard Coulomb type in 3D space, whereas for the case $r << R$, the potential decreases as $\sim 1/r^2$ as expected for 4D space with effective coupling strengthened by a factor of approximately $2\pi R$. In general, a coupling constant $g_{4+\delta}$ in $\delta = D - 4$ ED has mass dimension $-\delta/2$ and its strength is reduced effectively to $g_{\text{eff,4D}}$ in 4D like

$$g_{\text{eff,4D}}^2 \approx \frac{g_{4+\delta}^2}{V^\delta} = \frac{g_{4+\delta}^2}{(2\pi R)^\delta} \tag{6.16}$$

This formula is consistent with the effective gravitational strength obtained in Eq. (6.2).

$$G_N = G_N^{(4\text{D})} = \frac{1}{M_{\text{Pl}}^2} \sim \frac{1}{M_D^{\delta+2}}\frac{1}{R^\delta} \sim \frac{G_D}{V^\delta} \tag{6.17}$$

6.3
Chiral Fermions

Fermions in the ED is more complicated than scalars and vectors because of the presence of γ matrix and chirality [525–527]. One possible generalization of the 4D

---

7) Note $4\pi^2 r^2$ is the total surface area of a sphere in 4D space.

Dirac equation to 5D is

$$(i\partial_M \Gamma^M - M)\Psi(x, y) = 0 \tag{6.18}$$

where 4×4 Γ matrices satisfy the Clifford algebra

$$\begin{aligned}
&\left\{\Gamma^M, \Gamma^N\right\} = 2\eta^{\mathrm{MN}} \\
&\Gamma^\mu = \gamma^\mu, \quad \Gamma^5 = i\gamma^5 \\
&\gamma^\mu = \begin{bmatrix} 0 & \sigma^\mu \\ \overline{\sigma}^\mu & 0 \end{bmatrix}, \quad \gamma^5 = \begin{bmatrix} -\mathbf{1} & 0 \\ 0 & \mathbf{1} \end{bmatrix} \\
&\sigma^\mu = (1, \boldsymbol{\sigma}), \quad \overline{\sigma}^\mu = (1, -\boldsymbol{\sigma})
\end{aligned} \tag{6.19}$$

The fermion field on a circle (S^1) can be decomposed similarly to the scalar field with φ, ϕ in Eq. (6.7) replaced with $\tilde{\psi}$, ψ.

$$\begin{aligned}
\Psi(x, y) &= \frac{1}{\sqrt{2\pi R}} \sum_{n=-\infty}^{+\infty} \tilde{\psi}^{(n)}(x) e^{\frac{iny}{R}} \\
&= \frac{1}{\sqrt{2\pi R}} \psi^{(0)}(x) + \frac{1}{\sqrt{\pi R}} \sum_{n=1}^{\infty} \left\{ \psi^{(+n)}(x) \cos\frac{ny}{R} + \psi^{(-n)}(x) \sin\frac{ny}{R} \right\}
\end{aligned} \tag{6.20}$$

Carrying out the y-differentiation, the Dirac equation in 5D can be converted to that in 4D.

$$\left[i\partial_\mu \gamma^\mu - \left(M + i\frac{n}{R}\gamma^5 \right) \right] \psi^{(n)}(x) = 0 \tag{6.21}$$

Thus, from the 4D point of view, the Dirac spinors in 5D has a tower of particles with the mass $m_n = \sqrt{M^2 + (n/R)^2}$, as expected.[8] The mass spectrum of the fermion is illustrated in Figure 6.3a.

Compactification of the fifth dimension breaks Lorentz invariance. In 4D, γ^5 is a Lorentz-invariant operator, and has eigenvalue ± 1, a fact that is used to differentiate two kinds of fermions, namely the left-handed (LH) and the right-handed (RH) states.

$$\psi_L = \frac{1 - \gamma^5}{2}\psi, \quad \psi_R = \frac{1 + \gamma^5}{2}\psi \tag{6.22}$$

They are Lorentz invariants in 4D, although they mix by interactions. In the SM, the LH and RH particles have different quantum numbers and distinction of the chirality is essential. In 5D, γ^5 is on equal footing with other γ matrices. It is no longer Lorentz-invariant in 5D.[9] Both chiral states transform identically under some gauge symmetry one can think of. Hence, such a scenario cannot correspond to the SM in 4D where the LH and RH states transform differently as doublets

---

[8] The energy-momentum mass relation is obtained by applying $(i\partial_M \Gamma^M + M) = [i\partial_\mu \gamma^\mu + M - (in/R)\gamma^5]$ to Eq. (6.21).

[9] In 4D, the Lorentz boost operator is expressed as $\exp[-i\omega_{0\mu} S^{0\mu}]$ where $\omega_{0\mu} = (0, \eta_1, \eta_2, \eta_3)$ (η_i are rapidities) and $S^{0\mu} = (i/2)[\gamma^0, \gamma^\mu]$. Then, the 5D Lorentz boost along the fifth coordinate has the form $\exp[-i\omega_{05} S^{05}/2]$ (see Eq.(4.41) in [1]). The Lorentz operator no longer commutes with γ^5. In fact, it flips the chiral state.

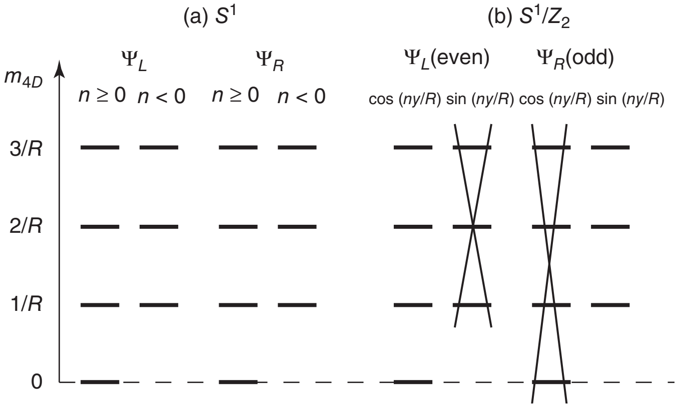


Figure 6.3 KK decomposition for a 5D fermion on a circle (S^1) (a) and on an orbifold (S^1/Z_2) (b) with even parity for Ψ_L.

and singlets in $SU(2)$. Compactification of the fifth dimension breaks the Lorentz invariance. Chiral theory can be obtained if one requires an additional symmetry, that is, invariance under the y-parity operation.

This can be seen as follows: Γ^5 anticommutes with all other Γ^Ms [Eq. (6.19)]. Then, by inspecting the Dirac equation Eq. (6.18), one sees that, if $\Psi(x, y)$ satisfies the 5D Dirac equation, so does $\Psi^P(x, y) \equiv i\Gamma^5\Psi(x, -y)$. Note that this is true only for massless fermions because the mass term violates the y-parity reversal invariance. It will be remedied in due course. One sees that $i\Gamma^5$ is a parity operator for the fifth coordinate y, and the 5D Dirac equation is invariant under the y-parity transformation.

$$\begin{aligned}\Psi_L(x, y) &\to \Psi_L^P \equiv i\Gamma^5\Psi_L(x, -y) = +\Psi_L(x, -y)\\ \Psi_R(x, y) &\to \Psi_R^P \equiv i\Gamma^5\Psi_R(x, -y) = -\Psi_R(x, -y)\end{aligned} \tag{6.23}$$

The essential point is that LH and RH have different y-parities. By requiring invariance under y-parity transformation $y \to -y$, only the zero and even mode ($\psi^{(0)}$ and $\psi^{(+n)}$) components in the KK tower of ψ_L and odd modes ($\psi^{(-n)}$) of ψ_R survive. That is, we eliminated (projected out) the zero-mode $\psi_R^{(0)}$ of the RH fermions. Thus, we succeeded in extracting a part of the chiral fermions.

The fact is, we have shown how to project out the RH fermions, but we have not shown how to extract the RH fermions. If an RH fermion has different quantum numbers from those of LH fermions, it is possible to choose the y parity such that ψ_R has zero modes. But before going to the RH fermions, we need to understand a little more what effect the y-parity requirement brings to the compactified space.

6.3.1
Orbifold S^1/Z_2

To obtain chiral fermions in 5D, it was necessary to require a symmetry for the y-parity transformation ($y \to -y$, which is referred to as the Z_2 *symmetry*) on a circle.

In other words, we required

$$f(y) = \pm f(-y) \qquad \text{for } y \in S^1 \tag{6.24}$$

The condition $y \in S^1$ requires $f(y + 2\pi R) = f(y)$, but otherwise any point is equal to any other point. Mathematically, a circle (S^1) is a smooth manifold as it has no special points. Contrary to compactification on S^1, the requirement of y-parity symmetry not only makes half of the circle equivalent to the other half but also creates two special points, $y = 0$ and $y = \pi R$. Any other points produces + or − parity by the operation but the two points are fixed. Topologically, a circle S^1 is equivalent to a line with both ends identified. But the requirement of additional symmetry Z_2 has reduced the physical or fundamental domain of S^1 to $0 \leq y \leq \pi R$ but the two ends are not identical (see Figure 6.4). It is referred to as the *orbifold* and is denoted as S^1/Z_2.

To find the constraints of orbifolding on the fields, let us consider a scalar field that is defined on a line segment $y \in [-\pi R, \pi R]$ and can be decomposed to even and odd functions:

$$\Phi^{(+)} = \frac{1}{\sqrt{2\pi R}}\phi^{(0)}(x) + \sqrt{\frac{1}{\pi R}}\sum_{n=1}^{\infty}\phi^{(+n)}(x)\cos\tfrac{ny}{R}$$

$$\Phi^{(-)} = \sqrt{\frac{1}{\pi R}}\sum_{n=1}^{\infty}\phi^{(-n)}(x)\sin\tfrac{ny}{R} \tag{6.25}$$

The orbifold projection requires $y \to -y$ to be a good symmetry, and the fields defined on $[-\pi R, \pi R]$ can be categorized as even or odd under $y \to -y$. The requirement is equivalent to imposing Neumann or Dirichlet boundary conditions at the fixed points.

$$\begin{aligned}&\text{even:}\quad \Phi^{(+)}(-y) = \Phi^{(+)}(y)\\ &\to \partial_5\Phi^{(+)}(y=0) = \partial_5\Phi^{(+)}(y=\pi R) = 0, \qquad \text{(Neumann BC)}\end{aligned} \tag{6.26a}$$

$$\begin{aligned}&\text{odd:}\quad \Phi^{(-)}(-y) = -\Phi^{(-)}(y)\\ &\to \quad \Phi^{(-)}(y=0) = \Phi^{(-)}(y=\pi R) = 0, \qquad \text{(Dirichlet BC)}\end{aligned} \tag{6.26b}$$

Thus, we have shown how to project out the RH fermions. The method to extract the RH fermions will be shown soon.

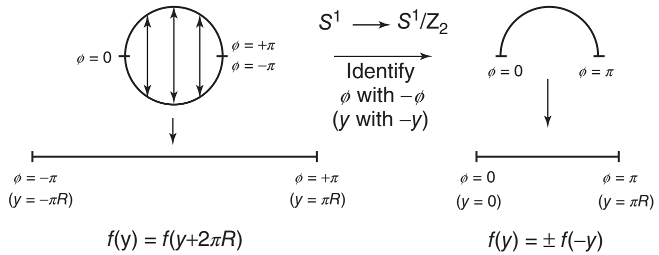


Figure 6.4 Orbifolding a circle (S^1) to an interval (S^1/Z_2).

Symmetry breaking by the boundary condition : The orbifolding we have just considered is equivalent to two requirements for a group element $\varphi(y)$ in $\mathscr{R}_1$, which takes any real number satisfying $-\infty < \varphi < +\infty$:

$$\varphi(y) \to \varphi'(y) \equiv S\varphi(y) = \varphi(y + 2\pi R) \tag{6.27a}$$

$$\varphi(y) \to \varphi'(y) \equiv Z\varphi(y) = \varphi(-y) = \pm\varphi(y) \tag{6.27b}$$

In field theory, one can relax the condition Eq. (6.27a) for the two identified points to represent the same physics. In general, if there is an extra global or local symmetry T in the Lagrangian, the transformed field does not have to be exactly equal, but merely equal up to the symmetry transformation, that is

$$\varphi(y + 2\pi R) = T\varphi(y) \tag{6.28}$$

The Lagrangian is bilinear in φ and invariant under the transformation Eq. (6.28) and describes the same physics. However, there must be a consistency condition between the two operations Z and T. To find it, let us apply the Z transformation first to an arbitrary point y and then a translation by $2\pi R$. The operation brings y to $2\pi(R - y)$. Another way to connect the two points is to apply the translation backward first by $y \to y - 2\pi R$ and then reflection around $y = 0$, which will also take the point to $2\pi R - y$. This means

$$ZT^{-1} = TZ, \qquad \to ZTZ = T^{-1} \tag{6.29}$$

Note $Z^2 = 1$, but $T^2 \neq 1$. However, for nontrivial T's ($T \neq 1$: T is referred to as the *Scherk-Schwartz twist*), we can define another operator Z_2' by

$$Z_2{}' \equiv TZ \qquad \to \qquad (Z_2')^2 = TZTZ = TT^{-1} = 1 \tag{6.30}$$

Therefore, if one has an extra global (or local) symmetry T, one can define another reflection symmetry by choosing a suitable transformation T.

Let us apply the above ansatz to obtain a zero-mode ψ_R. Suppose that there is a $SU(2)$ flavor symmetry $\Psi = (\psi_1, \psi_2)$.[10] Let us consider a symmetry operator $T \subset SU(2)$, $T^2 = 1$. An obvious choice is $T = \sigma_3$. By T operation

$$\Psi = \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} \quad \xrightarrow{T} \quad \begin{bmatrix} + & \\ & - \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} \tag{6.31}$$

By requiring $Z_2{}' = Z_2 T$ parity to be even instead of Z_2, we can have zero modes for both ψ_{1L} and ψ_{2R}. This is the promised extraction method of obtaining the zero mode of the RH fermions. It is straightforward to enlarge the symmetry to $SU(3)$ so that we can treat the t quark differently from the u, c quark. Notice, however, that by this orbifold boundary conditions, we have broken the $SU(2)$ symmetry to $U(1) \times U(1)$. This way of symmetry breaking by the boundary condition is referred to as the *Scherk–Schwartz mechanism.*[11]

It will be used later to obtain zero mode for the fifth component (A_5) of a 5D gauge field to identify it as the SM Higgs in an effort for gauge-Higgs unification.

---

[10] This is not the conventional isospin symmetry to work on u and d. Rather, it should be considered as the horizontal or family symmetry to differentiate u and c.

[11] It is obvious that this mechanism can never occur in the Minkowski or any spacetime that is homeomorphic (one-to-one correspondence) to $\mathscr{R}^n$.

6.3.2
Mass Generation and Localization

We saw that fermions can be made chiral by compactifying the fifth dimension on an orbifold, but only for the massless fermions in 4D. This is unsatisfactory because the SM has a highly hierarchical mass structure depending on the flavor. The reason we had to assume the massless fermion was due to noninvariance of the mass term for the $y \to -y$ symmetry. It is reminiscent of the SM where the fermions are assumed to be massless because of the chiral symmetry. The mass was generated by the chiral invariant Yukawa interaction with the Higgs. Therefore, we can try a similar remedy.

Let us introduce a y-reversal invariant scalar field with a potential such that it can acquire the VEV v spontaneously. For a while, we forget about the chiral structure and concentrate on the mass generation mechanism. We will come back soon to reintroduce the orbifold to use the mechanism in chiral theory.

We introduce a Yukawa interaction for the massless fermion and a scalar field equipped with a potential similar to that of the Higgs in 4D. Their action in 5D is given by

$$\begin{aligned} S &= S_F + S_H & \text{(6.32a)} \\ &= \int d^4x dy \left[\overline{\Psi} \left(i\Gamma^M \partial_M - h\Phi \right) \Psi + M_* \left\{ \frac{1}{2} \partial_M \Phi \partial^M \Phi - V(\Phi) \right\} \right] & \text{(6.32b)} \\ V(\Phi) &= \frac{\lambda}{4} (\Phi^2 - v^2)^2 & \text{(6.32c)} \end{aligned}$$

where we introduced the mass coefficient M_* conveniently so that the scalar field Φ has the dimension of mass. h is the Yukawa coupling strength.

The scalar field is constrained to satisfy

$$\Phi(x, y) \xrightarrow{y \to -y} -\Phi(x, -y) \tag{6.33}$$

so that the Yukawa interaction $\overline{\Psi}\Phi\Psi = (\overline{\Psi}_L \Psi_R + \overline{\Psi}_R \Psi_L)\Phi$ satisfies the y-parity symmetry.

The domain configuration of the scalar field The equation of motion for the field satisfies

$$\left[\partial_\mu \partial^\mu - \partial_y^2 \right] \Phi = -\lambda(\Phi^2 - v^2)\Phi \tag{6.34}$$

The potential $V(\Phi)$ has the shape depicted in Figure 6.5. For a potential with multi-vacuum like the one in Figure 6.5a, a special solution referred to as the *soliton* or the *kink* exists.[12] For Eq. (6.34), assuming $\Phi(x, y) = \Phi(y)$, the solution is

---

[12] The soliton appears in various disguises. In 1D, it is referred to as the *kink* or *string*, in 2D as *domain wall*, in 3D as *monopole*, and so on. In condensed matter physics, it is known as *topological defects*. Some examples are the Abrikosov vortex (quantized magnetic fluxes that penetrate superconductors of the second kind), domain walls between magnetic domains of differing orientation, and so on. See Chapter 7.1.2 for detailed discussions.

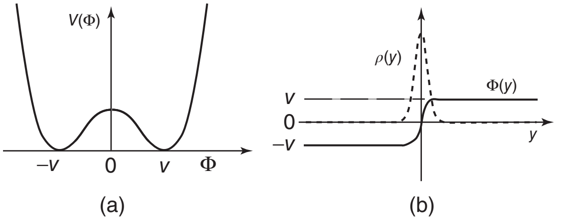


Figure 6.5 (a) Potential for the scalar field Φ. $\Phi = \pm v$ are the minimum of the potential. (b) For the soliton solution, $\Phi(\pm\infty) = \pm v$, which is referred to as the *domain (kink in 1D) configuration*. The energy of Φ, and hence that of Ψ, is localized near $y \approx 0$.

expressed as

$$\Phi(y) = \pm v \tanh\left(\sqrt{\frac{\lambda}{2}}v(y-y_0)\right) \quad \begin{cases} \xrightarrow{\lambda\to\infty} \pm v\varepsilon(y-y_0) \equiv \begin{cases} \pm v & (y > y_0) \\ \mp v & (y < y_0) \end{cases} \\ \xrightarrow{\lambda \ll 1/|v(y-y_0)|^2} \simeq \pm\sqrt{\frac{\lambda}{2}}v^2(y-y_0) \end{cases} \tag{6.35}$$

Notice that the VEV of the scalar field is constant in 4D but changes sign in y-space. When the solution is inserted back in Eq. (6.32), its energy density $\rho = 2V(\Phi)$ is localized around $y = y_0$ and is expressed as [see Eq. (7.5) and (7.9)]

$$\rho(y) = \frac{\frac{\lambda v^4}{2}}{\cosh^4\left(\sqrt{\frac{\lambda}{2}}v(y-y_0)\right)} \tag{6.36}$$

The solution for Φ and its energy density are plotted in Figure 6.5b for $y_0 = 0$. When the sign of Φ is flipped, it has a solution having the same energy density but with different topological boundary conditions. The two solutions are referred to as the *domain* (kink in 1D) and the *anti-domain* (anti-kink in 1D). As the requirement of the scalar field being odd under Z_2 parity forces $\Phi(0) = 0$, the scalar field has kink–anti-kink configuration at the fixed points. That is, they are flat except in the neighborhood of $|y| \lesssim 1/(\sqrt{\lambda/2}v)$ where $\Phi(y) \propto y$. In the limit $\lambda \to \infty$, the solution is a step function and is localized exactly at $y = 0$, that is, it is a physical example of a brane at $y = 0$ in 5D bulk. For a finite λ, the domain has a finite width $\sim 1/(\sqrt{2\lambda}v)$, which is referred to as a *fat brane* [528]. Notice that the scalar field acquires the VEV as a result of the spontaneous symmetry breaking, but unlike the Higgs in the SM, our solution Φ is not a constant but a function of y and has a localized energy density. $\Phi(y)$ is only asymptotically constant [$\Phi(y \to \pm\infty) = \pm v$ (or $\mp v$)] and changes sign for $y \gtrless 0$.

Domain-Coupled Fermions Given the background scalar configuration like one just stated, the fermion field Ψ satisfies the field equation

$$\left[i\Gamma^M\partial_M - \Phi(y)\right]\Psi(x,y) = \left[i\gamma^\mu\partial_\mu - \gamma^5\partial_y - \Phi(y)\right]\Psi(x,y) = 0 \tag{6.37}$$

We have set $h = 1$ for simplicity. We seek a solution given by

$$\Psi(x, y) = \sum_n \left[\psi_L^{(n)}(x) f_L^{(n)}(y) + \psi_R^{(n)}(x) f_R^{(n)}(y)\right] \tag{6.38}$$

where $\psi_{L(R)}^{(n)}(x)$ are solutions of the 4D Dirac equation and satisfy

$$\begin{aligned} i\gamma^\mu \partial_\mu \psi_L^{(n)}(x) &= m_n \psi_R^{(n)}(x), \qquad & i\gamma^\mu \partial_\mu \psi_R^{(n)}(x) &= m_n \psi_L^{(n)}(x) \\ \gamma^5 \psi_L^{(n)}(x) &= -\psi_L^{(n)}(x), & \gamma^5 \psi_R^{(n)}(x) &= \psi_R^{(n)}(x) \end{aligned} \tag{6.39}$$

Substituting Eqs. (6.38) and (6.39) into Eq. (6.37), we obtain the equations for $f_{L,R}^{(n)}(y)$.

$$\begin{aligned} [-\partial_y + \Phi(y)] f_L^{(n)}(y) &= m_n f_R^{(n)}(y) \\ [+\partial_y + \Phi(y)] f_R^{(n)}(y) &= m_n f_L^{(n)}(y) \end{aligned} \tag{6.40}$$

Now, let us define $Q_\pm$ operators by

$$Q_+ \equiv -\partial_y + \Phi(y), \qquad Q_- \equiv +\partial_y + \Phi(y) \tag{6.41}$$

They satisfy $Q_\pm^\dagger = Q_\mp$ and the Schrödinger equations

$$\begin{aligned} H_L f_L^{(n)} &\equiv Q_- Q_+ f_L^{(n)} = [-\partial_y^2 + \Phi(y)^2 + \partial_y \Phi(y)] f_L^{(n)} = m_n^2 f_L^{(n)} \\ H_R f_R^{(n)} &\equiv Q_+ Q_- f_R^{(n)} = [-\partial_y^2 + \Phi(y)^2 - \partial_y \Phi(y)] f_R^{(n)} = m_n^2 f_R^{(n)} \end{aligned} \tag{6.42}$$

Each $f_L^{(n)}, f_R^{(n)}$ forms an orthonormal set for the Hamiltonian $H_{L,R}$. Let us consider $\Phi(y) = -2\mu^2 y$, which is one of the approximate solutions for Eq. (6.34) in the neighborhood of $y \simeq 0$.[13] Then, the problem reduces to that of a simple harmonic oscillator. As $Q_\pm$ satisfies

$$[Q_+, Q_-] = -2\partial_y \Phi = 4\mu^2 \tag{6.43}$$

we see that

$$a^\dagger \equiv \frac{Q_-}{2\mu}, \qquad a \equiv \frac{Q_+}{2\mu} \tag{6.44}$$

are the creation and annihilation operators for the Hamiltonian H_L. The role is reversed for the Hamiltonian H_R with $\Phi = 2\mu^2 y$. For the ground state mode

$$\begin{aligned} a f_L^{(0)}(y) &= 0, \qquad \rightarrow \qquad \left(-\partial_y + \Phi(y)\right) f_L^{(0)}(y) = 0 \\ a^\dagger f_R^{(0)}(y) &= 0, \qquad \rightarrow \qquad \left(\ \ \partial_y + \Phi(y)\right) f_R^{(0)}(y) = 0 \end{aligned} \tag{6.45}$$

Thus,

$$f_L^{(0)}(y) \propto e^{+\int_0^y dy' \, \Phi(y')}, \qquad f_R^{(0)}(y) \propto e^{-\int_0^y dy' \, \Phi(y')} \tag{6.46}$$

For the wave function to be normalizable, we need $\Phi(\infty) < 0$ for $f_L^{(0)}(y)$ and $\Phi(+\infty) > 0$ for $f_R^{(0)}(y)$. Therefore, for the anti-domain configuration $\Phi(y) = -2\mu^2 y$, only $f_L^{(0)}(y)$ can be the solution. $f_R^{(0)}(y)$ is not allowed. In other words, once we choose one of the domain configurations, only one of f_L, f_R is allowed. It is well known in quantum mechanics that the wave functions for harmonic oscillators are given by

---

[13] It means that we adopt the anti-kink solution $\Phi(\pm\infty) = \mp v$ in Eq. (6.35). It is equivalent to setting $\lambda = 8$, $v = \mu$ in Eq. (6.35) and to use the approximation $\tanh x \simeq x$ for small x.

$$f_L^{(0)}(y) = \left(\frac{2\mu^2}{\pi}\right)^{\frac{1}{4}} \exp\left[-\mu^2 y^2\right], \qquad f_L^{(n)}(y) = \frac{(a^\dagger)^n}{\sqrt{n!}} f_L^{(0)}(y) \tag{6.47}$$

Exactly the same wave functions are obtained for $f_R^{(n)}(y)$ for the other choice of the domain configuration.

Chiral Fermions in the Brane What we have shown in the previous paragraph is that the fermions coupled with the scalar fields can be localized in the domain. The formalism we have developed is valid for y, which is allowed to go to $\pm\infty$. When we compactify the fifth dimension to place y on a circle (i.e., $-\pi R \le y \le \pi R$), the situation is different. Let us consider a limiting case ($\lambda \to \infty$) in which $\Phi(y) = -v\epsilon(y) = -M\epsilon(y)$ can be assumed. The eigen state equations for the fermion are now

$$\left[-\partial_y - M\epsilon(y)\right] f_L^{(n)}(y) = m_n f_R^{(n)}(y) \tag{6.48a}$$

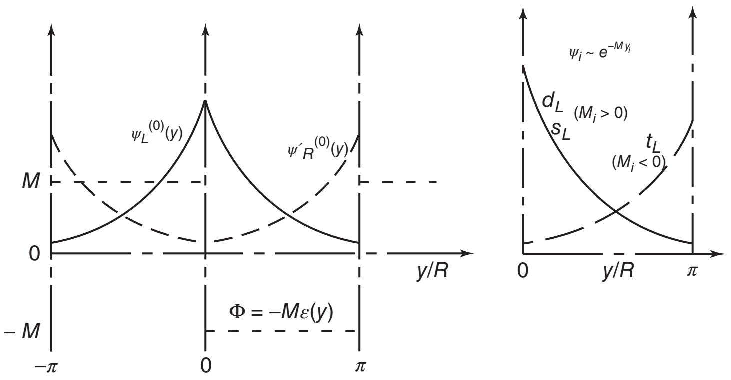
$$\left[\ \partial_y - M\epsilon(y)\right] f_R^{(n)}(y) = m_n f_L^{(n)}(y) \tag{6.48b}$$

Notice that $-M\epsilon(y)$ has its origin in the Yukawa interaction with the scalar field. It corresponds to the fermion's 5D mass but the 4D mass is given by $m_n = \sqrt{M^2 + (n/R)^2}$. As y does not go to $\pm\infty$, solutions exist for both $f_{L,R}^{(n)}(y)$. For $M = 0$, trigonometric (cos, sin) functions are the solutions, and we have the familiar KK tower given by Eq. (6.20). But, for $M \neq 0$, the factor $\epsilon(y)$ induces $\partial_y\Phi(y) \sim \delta(y)$, and the solutions to Eq. (6.42) are not the trigonometric functions. Let us concentrate, for simplicity, on the zero mode. The solution to f_L in Eq. (6.48) is given by

$$f_L^{(0)}(y) = \begin{cases} Ne^{-My} & (0 \le y \le \pi R) \\ Ne^{My} & (-\pi \le y \le 0) \end{cases} \tag{6.49a}$$

$$N(y > 0) = \left[\int_0^{\pi R} |e^{-My}|^2 dy\right]^{-\frac{1}{2}} = \left[\frac{2M}{1 - e^{-2M\pi R}}\right]^{\frac{1}{2}} \tag{6.49b}$$

which is depicted in Figure 6.6a. In other words, by introducing the step function $\epsilon(y)$, we have created discontinuities at $y = 0$ and $y = \pm\pi R$. One can see that Eqs. (6.48) and (6.49) satisfy the Z_2 symmetry. By orbifolding and requiring further that the wave functions be even under the Z_2 parity, the zero mode was constrained to contain dominant LH states. The RH state has solution $e^{M|y|}$, which does not vanish at $y = 0$. It contradicts the condition that the odd parity solution has to vanish at $y = 0$. Therefore, there are no RH fermion modes here. It is possible to interchange the role of Ψ_L and Ψ_R by requiring the opposite parity. In this case, there is no LH zero mode. Coexistence of both states is possible if the LH and RH



(a) (b)

Figure 6.6 (a) Profile of the scalar field (straight dashed line) for an anti-domain configuration at $y = 0$ and a domain at $y = \pi R$. The zero mode of the LH fermion is depicted by the solid line. If the RH fermion belongs to a different symmetry group (denoted as Ψ'_R), we can choose it to be an even function too. The dashed curve denotes its wave function. (b) Flavor hierarchy: SM particles are assumed to reside in the $y = \pi R$ brane. Each flavor quark has a different value of M. d_L, s_L have $M > 0$, but t_L has $M < 0$. With modest values $M_{d,s} \sim 2$, $M_t \sim -1$, the mass hierarchy can be explained.

fermions belong to different symmetry groups. We will explain how to construct the SM flavor hierarchy using this formalism.

6.3.3
Hierarchy

Having learned how to obtain chiral fermions, we now describe how they can be used to solve the big and the flavor hierarchy in the SM and whether they contradict some of the experimental facts, notably the proton stability.

Flavor Hierarchy Let us consider how the flavor (or mass) hierarchy in the SM can be achieved in the higher dimensions [527]. Take a case where the SM Higgs is localized at $y = \pi R$ brane. In this model, the fermions reside in the $y = 0$ brane but the SM is realized at $y = \pi R$ brane. That is, we only have a glimpse of particles residing in the other brane. The Yukawa interaction at $y = \pi R$ is described by the action

$$S_{5D}|_{y=\pi R} = \lambda_5 \int d^4x dy \delta(y - \pi R)\Phi(y)\overline{\Psi}_L(y)\Psi'_R(y) \tag{6.50}$$

where Ψ and Ψ' are two different 5D fermion fields that belong to SU(2) doublets and singlets, with M, M' being their 5D masses, respectively. Here, both $\Psi_L(y) \sim e^{-My}$ and $\Psi'_R(y) \sim e^{M'y}$ are chosen to be even under Z_2 so that they give the LH and RH zero modes, respectively. The opposite signs for Ψ_L and Ψ'_R follow from Eq. (6.48a) and Eq. (6.48b). Ψ_R and Ψ'_L are odd and vanish at $y = 0$, πR and do not couple to the scalar field. Hence, they give no contribution to the mass. The

4D action is given by integrating over y.

$$S_{4\text{Deff}} = \lambda_4 \int d^4x\, \Phi \overline{\Psi}_L \Psi'_R \qquad \lambda_4 \simeq \lambda_5 e^{(M'-M)\pi R} \tag{6.51}$$

Let us assume, for simplicity, that d, s and t quarks have the same λ_5 and also $M = -M'$ for each quark. Then, for the choice of $\Delta M \pi R \sim -2$ or 5 (e.g., $M_d = 3/\pi R$, $M_s = 1/\pi R$, $M_t = -2/\pi R$)

$$\begin{aligned} \frac{m_d}{m_s} &\sim \frac{e^{(M'_d - M_d)\pi R}}{e^{(M'_s - M_s)\pi R}} = e^{-2(M_d - M_s)\pi R} \equiv e^{2\Delta M \pi R} = e^{-4} \sim \frac{1}{100} \\ \frac{m_d}{m_t} &\sim e^{-2(M_d - M_t)} = e^{-10} \sim 5 \times 10^{-5} \end{aligned} \tag{6.52}$$

which reproduces approximately the right observed values. Sign flip for the top means that we consider the anti-domain configuration or equivalently, the top, unlike other particles, primarily is assumed to reside in the $y = \pi R$ plane (see Figure 6.6b). The crucial point is that we did not invoke any large hierarchies in the fundamental parameters (λ_5 or M), yet we could obtain the large mass/flavor hierarchy observed in the SM.

Big Hierarchy The renormalization group equation (RGE) to describe running coupling constants was discussed in Section 3.6. The RGE for SUSY with β functions given by Eq. (3.89) is expressed as

$$\begin{aligned} &\frac{1}{\alpha_i(\mu)} = \frac{1}{\alpha_i(\mu_0)} + \frac{\beta_i}{2\pi} \ln \frac{\mu}{\mu_0}, \qquad \alpha_i = \frac{g_i^2}{4\pi} \\ &(\beta_1, \beta_2, \beta_3) = \left(-\frac{33}{5}, -1, 3\right) \end{aligned} \tag{6.53}$$

Setting $\mu_0 = m_Z$, $\mu = M_{\text{GUT}}$ and plugging the experimental values for α_is, one gets the celebrated gauge unification (see Figure 3.11)

$$\alpha_1(M_{\text{GUT}}) = \alpha_2(M_{\text{GUT}}) = \alpha_3(M_{\text{GUT}}) \approx \frac{1}{25} \tag{6.54}$$

at $M_{\text{GUT}} \simeq 2 \times 10^{16}$ GeV. The Planck scale ($M_{\text{Pl}} = 1.22 \times 10^{19}$ GeV is slightly above M_{GUT}. This large discrepancy between the weak scale $G_F^{-\frac{1}{2}} \sim 250$ GeV and the GUT-Planck scale is the "gauge or big hierarchy problem."

What will happen to the above relations in the ED scenario? The beta functions are contributions of the higher order loop integrals of all the MSSM particles having mass below μ. Let us assume that all the gauge particles reside in the ED. In this case, contributions of the KK towers have to be included. We assume that the KK masses appear above $\mu_0 \equiv 1/R$. Notice, however, that higher dimensional gauge theories are nonrenormalizable because the coupling constants have the inverse mass dimension. So we will not be talking of running coupling constants but rather corrections to them with infrared cutoff μ_0 and ultraviolet cutoff Λ. Below $\mu < \mu_0$, the physics is effectively described by that of MSSM. The Λ should be regarded as a boundary beyond which a new physics (correct ED theory) will appear. Thus, our treatment should be regarded as an effective theory valid only below Λ. With those caveats in mind, we obtain the following RGE after including contributions of the KK towers [529, 530]:

$$\frac{1}{\alpha_i(\Lambda)} = \frac{1}{\alpha_i(m_Z)} + \frac{\beta_i}{2\pi}\ln\frac{\Lambda}{m_Z} - \frac{\tilde{\beta}_i}{2\pi}\ln\frac{\Lambda}{\mu_0} + \frac{\tilde{\beta}_i X_\delta}{2\pi\delta}\left[\left(\frac{\Lambda}{\mu_0}\right)^\delta - 1\right]$$

$$\delta = D - 4, \qquad X_\delta = \frac{\pi^{\frac{\delta}{2}}}{\Gamma\left(1+\frac{\delta}{2}\right)} \tag{6.55}$$

$$(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3) = \left(-\frac{3}{5}, 3, 6\right) - n_{\text{exg}}(4, 4, 4)$$

where n_{exg} is the number of extra generations.

The first and the second terms are those of usual MSSM given in Eq. (6.53), which in our language correspond to the zero mode states. The new beta functions $\tilde{\beta}_i$ are due to KK modes, with each contribution matching to those of zero mode states. Note that the fourth term on the right-hand side, which describes contributions of infinite numbers of KK mode particles, has power dependence in Λ. The presence of EDs has changed the logarithmic evolution into a power evolution! The third term is the $\delta = D - 4 \to 0$ limit of the fourth term, which is subtracted to give the correct $\delta \to 0$ expression.

Numerical results of Eq. (6.55) are given in Figure 6.7.

One sees that all the three forces are unified at $\Lambda \simeq 20$ TeV. This is a reduction of the unification scale by order 10 compared to the conventional SUSY GUT

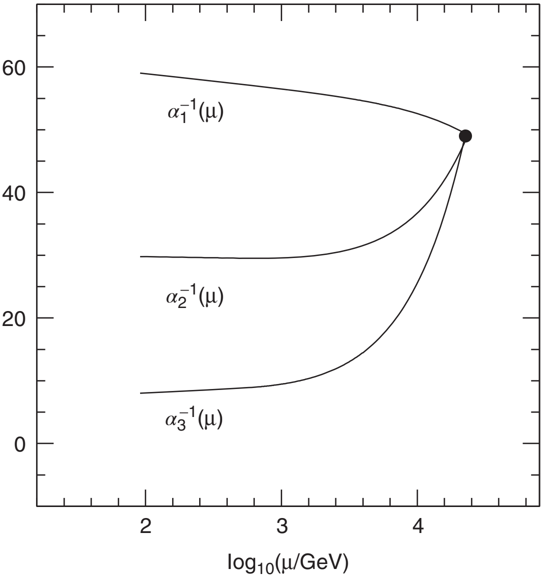


Figure 6.7 Unification of gauge couplings in the presence of extra spacetime dimensions at $\Lambda \simeq 20$ TeV. The figure is obtained for parameter settings $\mu_0 = 1/R = 1$ TeV, $\delta = 1$, $n_{\text{exg}} = 0$. [529]

unification scale. The unification scale Λ may differ depending on parameters, but it is worthwhile to note that the unification is a universal feature independent of the number of EDs (δ) and parameters such as μ_0, n_{exg}. The big hierarchy problem has disappeared in the ED theory! Note that this is achieved for the flat ED scenario. Another possibility to use warped space will be given later in the discussion of gravity.

Proton decay: Perhaps the most immediate question that follows from the reduction of the big hierarchy is the proton lifetime. Experimentally, it is known to have a lifetime of at least $\sim 10^{33}$ years. Traditional GUT theories achieve it by attaching a large mass to the gauge particles (denoted X, Y) or the GUT Higgs that are responsible for the baryon-number-violating processes [see Eq. (3.57)].

$$\tau_p = \tau_\mu \left(\frac{m_\mu}{m_p}\right)^5 \left(\frac{M_X}{m_W}\right)^4 \frac{BR(p \to e^+\pi^0)}{F} \sim \frac{M_X^4}{\alpha_5^2 m_p^5} \tag{6.56}$$

where $F \sim O(1)$ and $\alpha_5 \simeq \alpha_i(M_{\text{GUT}}) \sim 1/25$. For $M_X \sim M_{\text{GUT}} \sim 10^{16}$ GeV, it is consistent with observations. Obviously, it breaks down with small M_X as required by the ED scenario.

There are several mechanisms to achieve the requirement in the ED theory. Let us denote the proton-decay-mediating particle as Ψ_X. As it is not yet observed, it should not have zero mode in the ED theory, which means that it is odd under Z_2 symmetry. Then, its wave function in the higher dimension should behave like $\sim \sin(ny/R)$. If the fermions are localized at $y = 0$, πR, they have no interaction with Ψ_X and hence the proton is stable. If the fermions are not completely localized at the fixed points, the overlap of the wave functions with Ψ_X is naturally small. Notice that the mechanism does not depend on the mass of Ψ_X but rather on the inherent higher dimensional symmetry.

6.3.4
Split Fermion Scenario

Here we introduce yet another mechanism for explaining the flavor hierarchy. It uses a notion that the brane may have some finite width along the extra (i.e., along y) dimension (fat brane). This is the case for finite λ of the domain solution of the scalar fields given by Eq. (6.35). In this case, it is possible to localize the quarks and leptons at different points within the domain. This is a scenario referred to as *split fermions* [531]. Consider the domain-coupled fermions given by Eq. (6.37) and assume that the fermion "i" has intrinsic mass m_i. The action is written as

$$S = \int d^4x dy \sum_i \overline{\Psi}_i \left[i\Gamma^M \partial_M + \lambda_i \Phi(y) - m_i \right] \Psi_i \tag{6.57}$$

The field equation naturally takes the domain-coupled fermion fields to a point $\lambda_i \Phi(y) - m_i = 0$. That is, fermions with different m_i's are localized at different locations in y (see Figure 6.8).

What effects will the split locations have?

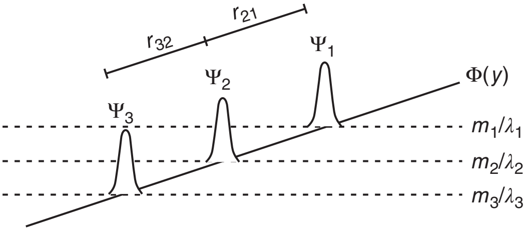


Figure 6.8 Split fermion: chiral fermions localized at different places in extra dimensions by a domain wall.

Flavor hierarchy: We already discussed one model to possibly explain the mass hierarchy [see Eq. (6.52)]. Here we present another model using the split fermion scenario. Let us assume that two fermions are separate in y position but within a fat brane. We consider the following action for the lepton fields:

$$
\begin{aligned}
S = \int d^4x dy \Big[& \overline{L}\Big(-\gamma^5 \partial_y - \Phi(y) \Big) L + \overline{E}^c \Big(-\gamma^5 \partial_y - \Phi(y) - m \Big) E^c \\
& - \Big(\kappa H L^T C_5 E^c + h.c. \Big) \Big] \; ^{14)}
\end{aligned}
\tag{6.58}
$$

where L is a doublet and E^c is a singlet member of the lepton fields. $\Phi \sim -2\mu^2 y$ in the vicinity of $y = 0$. Using the wave functions given by Eq. (6.47), the zero mode of the L field $l^{(0)}$ is localized at $y = 0$ and that of the E^c field $e^{c(0)}$ at $y = m/(2\mu^2) \equiv r$. The Yukawa coupling of the zero mode is then given by

$$
\begin{aligned}
S_{\text{Yukawa}} &= -\kappa \int d^4x\, h(x) l(x) e^c(x) \int dy f_l^{(0)}(y) f_{e^c}^{(0)}(y) \\
&= -\varepsilon\kappa \int d^4x\, h(x) l(x) e^c(x) \\
\varepsilon &= \int dy f_l^{(0)}(y) f_{e^c}^{(0)}(y) = \sqrt{\frac{2\mu^2}{\pi}} \int dy\, e^{-\mu^2 y^2} e^{-\mu^2 (y-r)^2} = e^{-\frac{\mu^2 r^2}{2}}
\end{aligned}
\tag{6.59}
$$

The Yukawa coupling is exponentially suppressed if the distance between the zero modes r is larger than $1/\mu$.

Proton decay : The proton decay operator in 5D is expressed as

$$
\begin{aligned}
S_{p\text{-decay}} &\sim \frac{g_5^2}{M_{\Psi_X}^2} \int d^4x dy \left(Q^T C_5 L \right)^\dagger \left(U^{cT} C_5 D^c \right) \sim \varepsilon' \frac{g_4^2}{m_W^2} \int d^4x \left(ql \right)^\dagger \left(u^c d^c \right) \\
\varepsilon' &= 2\pi R \left(\frac{m_W}{M_{\Psi_X}} \right)^2 \left(\frac{2\mu^2}{\pi} \right) \int dy \left(e^{-\mu^2 y^2} \right)^3 e^{-\mu^2 (y-r)^2} \sim e^{-\frac{3}{4}\mu^2 r^2}
\end{aligned}
\tag{6.60}
$$

14) $C_5 \equiv \Gamma^0 \Gamma^2 \Gamma^5$ is a charge conjugation operator in 5D. The two Lorentz-invariant bilinears in 5D are given by $\overline{\Psi}_1 \Psi_2$ and $\Psi_1^T C_5 \Psi_2$, which correspond to Dirac and Majorana mass terms in 4D.

where Ψ_X is the gauge particle to mediate the proton decay and we have put $\mu \sim M_{\Psi_X} \sim 1/\pi R$. Here, we have assumed almost degenerate quarks that are separate from the lepton. For $\mu r \approx 10$, we obtain $|\varepsilon'|^2 \sim 10^{-33}$ which provides sufficient suppression for the proton decay rate.

6.4
Gauge Field in ED

6.4.1
Action in 4D

We now turn to gauge interactions in the higher dimension. We first consider how a gauge field in 5D looks like in 4D. The 5D action of a $U(1)$ gauge field $A_M(x, y)$ can be expressed as

$$S^{(5\mathrm{D})}_{\mathrm{gauge}} = \int d^4x \int_0^{2\pi R} dy\, \mathscr{L}^{(5\mathrm{D})}_{\mathrm{gauge}} \tag{6.61a}$$

$$\mathscr{L}^{(5\mathrm{D})}_{\mathrm{gauge}} = -\frac{1}{4}F_{MN}F^{MN} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}F_{\mu 5}F^{\mu}_{5} \tag{6.61b}$$

Applying the same KK tower expansion as given by Eq. (6.7)

$$A_M(x, y) = \frac{1}{\sqrt{2\pi R}} A^{(0)}_M(x) + \frac{1}{\sqrt{\pi R}} \left[\sum_{n=1}^{+\infty} A^{(+n)}_M \cos(ny/R) + A^{(-n)}_M \sin(ny/R) \right] \tag{6.62}$$

and inserting the above expansion formula in the Lagrangian in Eq. (6.61), we obtain the effective 4D action given by

$$\begin{aligned} S^{(4\mathrm{D,eff})}_{\mathrm{gauge}} = \int d^4x \Bigg[& -\frac{1}{4}(F^{(0)}_{\mu\nu})^2 + \frac{1}{2}(\partial_\mu A^{(0)}_5)^2 \\ & + \sum_{\pm n=1}^{\infty} \left\{ -\frac{1}{4}\left(F^{(\pm n)}_{\mu\nu}\right)^2 + \frac{1}{2}m_n^2 \left(A^{(\pm n)}_\mu \pm \frac{1}{m_n}\partial_\mu A^{(\mp n)}_5 \right)^2 \right\} \Bigg] \end{aligned} \tag{6.63}$$

where we have set $m_n = n/R$. By applying a gauge transformation

$$\begin{aligned} & A_M(x, y) \to A'_M(x{\cdot}y) = A_M(x, y) + \partial_M \theta(x, y) \\ & \theta(x, y) = \frac{1}{\sqrt{2\pi R}}\theta^{(0)}(x) + \frac{1}{\sqrt{\pi R}} \sum_{n=1}^{\infty} \left[\theta^{(+n)}(x)\cos\left(\frac{ny}{R}\right) + \theta^{(-n)}(x)\sin\left(\frac{ny}{R}\right) \right] \end{aligned} \tag{6.64}$$

It is possible to choose "almost axial" gauge where A_5 has no y dependence

$$A_5(x, y) = A^{(0)}_5(x) \tag{6.65}$$

Then $A_5^{(\mp n)}$ vanishes and Eq. (6.63) becomes

$$S_{\text{gauge}}^{(4\text{D})} = \int d^4x \Bigg[-\frac{1}{4} F_{\mu\nu}^{(0)} F^{(0)\mu\nu} + \frac{1}{2} \partial_\mu A_5^{(0)} \partial^\mu A_5^{(0)} + \sum_{\pm n=1}^{\infty} \left(-\frac{1}{4} F_{\mu\nu}^{(\pm n)} F^{(\pm n)\mu\nu} + \frac{m_n^2}{2} A_\mu^{(\pm n)} A^{\mu(\pm n)} \right) \Bigg] \tag{6.66}$$

This is an action for a gauge field $[F_{\mu\nu}^{(0)}(x)]$ and a massless real scalar field $(A_5^{(0)}(x))$ and a series of massive vector bosons $(A_\mu^{(\pm n)})$.

It is interesting to note that the KK tower of the gauge fields have acquired mass by eating the KK tower of A_5. That is, the mass was generated by the spontaneous symmetry breaking of $U(1)$ in the fifth dimension.[15] The symmetry is not broken in 4D and the zero mode gauge field remains massless.

An important consequence of the gauge field going into 5D is that it has produced an extra scalar field. In general, lower spin states appear automatically by going into the ED. For instance, the spin-2 graviton produces a spin-1 massless vector field. This was the motivation of the original KK theory. By identifying the massless vector field with the electromagnetic gauge field, they tried to unify the gravity and the electromagnetic force.

From modern field theoretical point of view, the additional scalar field presents an interesting possibility. Just as the KK tower of the fifth component $A_5^{(n)}$ was eaten to give mass to the KK tower of the gauge fields, one suspects that a possibility of the additional zero-mode scalar may be eaten by the zero mode gauge field: in other words, a possibility of $A_5^{(0)}$ being the Higgs field. This is referred to as the "*gauge-Higgs unification*". We recall that a major motivation in introducing the supersymmetry was to control the divergence problem in the Higgs sector. As A_5 is a part of the gauge field, it has a built-in mechanism to suppress the divergence and provides an alternative to solve the hierarchy problem without supersymmetry.

6.4.2
Coupling Strength

By going into 5D, the gauge field has acquired mass dimension 3/2, and hence the gauge coupling has mass dimension −1/2. Expanding A_M in KK levels, and extracting its $n = 0$ component, the covariant derivative is expressed as

$$D_\mu = \partial_\mu + ig_5 A_\mu = \partial_\mu + ig_5 \frac{1}{\sqrt{2\pi R}} A_\mu^{(0)} + \cdots \tag{6.67}$$

We see

$$g_4 = \frac{g_5}{\sqrt{2\pi R}} \tag{6.68}$$

---

[15] Notice that, for non-Abelian gauge group, the scalar field A_5 is in the same adjoint representation as rest of the gauge field.

which reconfirms the relation of the coupling constants Eq. (6.13) in 4D and 5D that we derived for the scalar.

6.4.3
Gauge–Higgs Unification

The Higgs is at the core of the SM. It has provided the mass generation mechanism compatible with the gauge and chiral symmetry that govern the SM particles. However, it has its own problems. The problem lies in the very existence of the Higgs itself and arbitrariness of its properties where no underlying principle or symmetry is apparent. First, its radiative corrections diverge quadratically. There is no protecting symmetry in the SM to place the divergence under control (see discussions in Section 1.4). A possible remedy may be provided by the supersymmetry, and the supersymmetric generalization of the SM was already discussed in the previous chapter. Second, the Yukawa coupling constants are quite arbitrary, and as many parameters as the number of flavors exist.

The gauge–Higgs unification tries to solve these problems by regarding the Higgs as the fifth component of the gauge field in the 5D theory [532–535]. We have seen in the previous section that the zero mode of $A_5(x, y)$ is a scalar in 4D and that $n \neq 0$ KK modes of the gauge particles become massive by eating $A_5^{(n)}$ $(n \neq 0)$. It is natural to think that the zero mode of the gauge particles also become massive by eating $A_5^{(0)}$.

If the size of the fifth dimension R is infinite, the gauge symmetry would guarantee vanishing of the mass correction $(\delta m_A^2 = 0)$ to all orders. However, the scalar $A_5^{(0)}$ has no potential in 5D, as it is a part of a 5D gauge field. This is the major problem of the gauge–Higgs unification scenario. So we need to devise a way to generate a mass (quadratic) term and a self interacting potential (quartic and higher order) terms. It can be achieved by the Coleman-Weinberg mechanism [6], that is, generating higher order terms radiatively.

The Hosotani Mechanism If A_5 is to behave like the Higgs in SM, it has to acquire a VEV. Under normal circumstances, however, the VEV can easily be removed by a gauge transformation without changing the physics contents. We do not want this happen here. Therefore, we notice that the local gauge symmetry also requires simultaneous transformation of the fermion field. The gauge symmetry (in the fifth dimension) means that the Lagrangian is kept invariant under a set of transformations:

$$\psi(x, y) \to \psi'(x, y) = U\psi(x, y), \qquad U = e^{-i\alpha(y)} \subset U(1) \tag{6.69a}$$

$$A_5 \to A_5' = A_5 + \frac{1}{g}\partial_y \alpha(y) \tag{6.69b}$$

We decompose the gauge field into the constant background field and quantum fluctuation around the background:

$$A_5(x, y) = <A_5(x, y)> + [A_5(x, y) - <A_5(x, y)>] \equiv A_5^c + A_5^q(x, y) \tag{6.70}$$

The solution to the Dirac equation in the presence of the background field A_5^c is given by

$$\begin{aligned} \psi(x,y) \to \psi'(x,y) &= \psi(x)\, e^{-ie\int_{y_0}^{y} dy A_5^c} = \psi(x)\exp\left[-i\frac{\Theta_{AB}}{2\pi R}(y-y_0)\right] \\ \Theta_{AB} &\equiv e\oint A_5^c\, dy \end{aligned} \tag{6.71}$$

Rewriting ψ' as ψ, the periodic condition on 5D circle S^1 becomes

$$\psi(x, y+2\pi R) = \psi(x,y)\, e^{-i\Theta_{AB}} \tag{6.72}$$

The exponential factor is known as the *Aharonov–Bohm phase* (aka Wilson phase). In quantum mechanics, it represents a magnetic flux[16] inside a closed path, a physical quantity that is observable and gauge-invariant (see arguments in Section 18.3 of [1]). In 5D, it is not a magnetic flux but some real being analogous to it. As a result of its presence, the path around the circle (i.e., space spanned by y) is no longer simply connected. In other words, the nonzero value of the phase integral arises as a result of the compact space S^1 being not simply connected (see arguments in Section 7.1.3). It is physical and not reducible despite the vanishing of its corresponding field strength. Equation (6.71) means that the fifth component of the momentum is given by

$$p_y\psi(x,y) = -i\frac{\partial\psi}{\partial y}(x,y) = \frac{1}{R}\left(n - \frac{\Theta_{AB}}{2\pi}\right)\psi(x,y) \tag{6.74}$$

That is, the zero mode of the fermion has acquired mass $m_0 = \Theta_{AB}/2\pi R$. Notice that the $U(1)$ symmetry in 4D is not broken here. The mass was generated by the non-simple-connectedness of S^1.

The Aharonov–Bohm phase in the field theory is not generated by an external source such as an external magnetic field in quantum mechanics. It is a part of the gauge field, which is a dynamical object as we will show in the following. It is a VEV of the A_5 (integrated over a closed path) whose value is determined dynamically. Notice that the scalar field A_5 is under the influence of the Coleman-Weinberg potential that we denote as $V(A_5)$. This is because there are interactions between the scalar field A_5 and the fermion fields, and the field A_5 can acquire its self (potential)

---

[16]

$$\Theta_{AB} = \frac{e}{\hbar c}\oint_C d\mathbf{x}\cdot\boldsymbol{A} = e\int_S \boldsymbol{B}\cdot d\boldsymbol{S} = e\Phi \tag{6.73}$$

energy radiatively through fermion loops.[17] If the mass term (quadratic term of the Taylor-expanded potential) for $\Theta_{AB} = 0$ is negative, the vacuum is unstable, hence the symmetry is broken spontaneously and a finite value of Θ_{AB} is realized. The stable point of the vacuum and the mass of the field A_5 are determined by the condition

$$\frac{\partial V(A_5^c)}{\partial A_5^c} = 0, \qquad \frac{\partial^2 V(A_5^c)}{\partial (A_5^c)^2} = m_H^2 > 0, \qquad \Theta_{AB} = 2\pi R g A_5^c \tag{6.76}$$

Thus, we have a dynamical mass-generating mechanism through the Aharonov–Bohm effect. This is referred to as the *Hosotani mechanism* [538–540]. Here, the Higgs particle is a quantum excitation of the Aharonov–Bohm phase.

For the A_5 to be eligible as the SM Higgs, one needs a little more improvement. First, presence of the chiral fermions requires Z_2 symmetry and change of the compactified space from a sphere S^1 to an orbifold S^1/Z_2. Unfortunately, A_5 is odd under Z_2 operation; hence, the zero mode disappears and we cannot have massless A_5 in 4D to be eaten by other gauge fields. This is because A_M is a 5D vector, which transforms like $x^M = (x^\mu, y)$. The transformation $y \to -y$ also changes $A_5 \to -A_5$. Second, the gauge field belongs to the adjoint representation of the symmetry group. In the SM, the gauge symmetry is $SU(2) \times U(1)$. The SM gauge fields constitute an isospin triplet, whereas the Higgs is a doublet.

The two problems can be solved by enlarging the symmetry, for instance, to $SU(3)$ and breaking it to $SU(2) \times U(1)$ [541, 542]. It is achieved by requiring $Z_2' = Z_2 P$ symmetry, which is a combination of $y \to -y$ and a special $SU(3)$ transformation

$$\Psi = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix} \xrightarrow{P} \begin{bmatrix} + & & \\ & + & \\ & & - \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix} \tag{6.77}$$

Under this P transformation, the adjoint representation transforms like $A_\mu^{ij} \sim \psi_i \overline{\psi}_j - (1/3)\mathrm{Tr}[\Psi\overline{\Psi}]$ $(i, j = 1 - 3)$.

$$\left[\begin{array}{c|c} A_M^{ij} & A_M^{i3} \\ \hline A_M^{3j} & A_M^{33} \end{array}\right]_{i,j=1,2} \xrightarrow{P} \left[\begin{array}{c|c} + & - \\ \hline - & + \end{array}\right] \tag{6.78}$$

---

[17] An explicit expression of $V(A_5^c)$ for the simple case of $M^4 \times S^1$ (i.e., the one ED is compactified to a circle) is given by setting $D = 4$ in the following formula [536, 537].

$$V(A_5^c) = -\frac{(2\pi R)^{-4}}{2^{\frac{(D-1)}{2}} \pi^{\frac{D}{2}} \Gamma\left(\frac{D}{2}\right)} \int_0^\infty ds s^{D-1} \ln\left[1 - 2\cos\Theta_{AB}\, e^{-\sqrt{s^2+(2\pi Rm)^2}} + e^{-2\sqrt{s^2+(2\pi Rm)^2}}\right] \tag{6.75}$$

where D is the dimension of the extended Minkowski space and m is the fermion mass. Expanded in powers of A_5^c, Eq. (6.75) has negative quadratic and positive quartic coefficients as desired. A stable vacuum is realized for $\Theta_{AB} = \pi$.

By Z_2 (y-reversal), P, and $Z_2' = Z_2 P$ transformations, the gauge fields in the adjoint representation and the fermions in the fundamental representation transform like

	A_μ^{ij}	A_μ^{i3}/A_μ^{3j}	A_μ^{33}	A_5^{ij}	A_5^{i3}/A_5^{3j}	A_5^{33}	$(\psi_{1\,L}, \psi_{2\,L})$	$\psi_{3\,R}$
Z_2	$+$	$+$	$+$	$-$	$-$	$-$	$+$	$-$
P	$+$	$-$	$+$	$+$	$-$	$+$	$+$	$-$
Z_2'	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$+$

(6.79)

By asking Z_2' symmetry, we have obtained zero modes for A_μ^{ij}, A_5^{i3} (or A_5^{3j}). They are exactly the needed fields in the SM. The fermions that contain zero modes are reduced to $(\psi_{1\,L}, \psi_{2\,L}) \subset SU(2)$ and $\psi_{3\,R}$, which belongs to a singlet. Thus, by orbifolding with Z_2' symmetry, the symmetry $SU(3)$ is broken to $SU(2) \times U(1)$. Notice that A_μ^{ij} are gauge bosons in 4D corresponding to the gauge group $SU(2)$, and A_5^{i3}/A_5^{3j} are massless doublet scalars on the brane. When the latter acquires a VEV, the gauge group $SU(2) \times U_Y(1)$ further breaks down spontaneously to $U_{\text{EM}}(1)$.

We have achieved the gauge–Higgs unification. Starting from gauge fields and fermion fields only in 5D, we have succeeded in reproducing the fields needed in SM, which are doublet fermions, Higgs, a singlet fermion, and a gauge triplet. The Higgs is now protected with the gauge symmetry, possibly providing an alternative scenario to the supersymmetry for solving the hierarchy problem.

Having said that much, we mention that it remains to construct a realistic model. For some examples, we refer to [535, 543–545].

6.5
Gravitational Field

Here we consider a case that the gravitational field resides in the bulk but SM particles are confined to the brane. When gravity is included, we have to rely on a general coordinate framework (x^μ) rather than simple Minkowski coordinates [546, 547]. As gravity is the dynamics of spacetime, one has to start from defining the metric of the spacetime. In 4D, it is defined by

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu = \left(\eta_{\mu\nu} + \tilde{h}_{\mu\nu}\right) dx^\mu dx^\nu \equiv \left(\eta_{\mu\nu} + \frac{h_{\mu\nu}}{2\overline{M}_{Pl}}\right) dx^\mu dx^\nu \tag{6.80}$$

where $\overline{M}_{\text{Pl}} = 1/\sqrt{8\pi G_N}$ is the reduced Planck mass.[18] The factor 1/2 is conventional. For vacuum, $g_{\mu\nu} = \eta_{\mu\nu} = \text{diag}(+, -, -, -)$, which defines the Minkowski spacetime. It is the background spacetime in which dynamics of the matter work. When one wants to talk of gravitational interactions, one turns on the gravity field. If gravity is treated quantum mechanically, the Minkowski metric is the VEV

---

[18] The reduced Planck mass $\overline{M}_{\text{Pl}}$ is denoted as $\tilde{M}_{\text{Pl}}$ in other chapters. We did not unify the notations, as the field is changing fast and $\overline{M}_{\text{Pl}}$ is conventionally used in the discussion of ED. From now on, we sometimes use the word "Planck mass" to mean the reduced mass depending on the situation.

($\eta_{\mu\nu} =<g_{\mu\nu}>$) and $h_{\mu\nu}$ represents the graviton. It is well known that linear approximations work excellently in general relativity, that is, the higher order effect of $h_{\mu\nu}$ is negligible. Notice, however, the notion may not be true for the graviton in the ED.

The Einstein equation of general relativity defines the 4D spacetime structure in terms of the matter field. To derive it in the Lagrangian formalism, one starts from the action

$$\begin{aligned} S &= \int d^4x \sqrt{-g}\,(\mathscr{L}_g + \mathscr{L}_m) \\ \mathscr{L}_g &= \frac{R^{(4)}}{16\pi G_N} = \frac{\overline{M}_{\rm Pl}^2}{2} R^{(4)}, \qquad g = \det g_{\mu\nu} \end{aligned} \tag{6.81}$$

where $R^{(4)}$ is the Ricci scalar curvature in 4D obtained by contracting the Ricci tensor $R^{(4)}_{\mu\nu}$ (see Appendix Eq. (E.17)). $\mathscr{L}_m$ is the matter field Lagrangian density. Constructing the usual Euler–Lagrange equation of motion as a function of the metric tensor $g_{\mu\nu}$, one obtains the Einstein equation.

$$\mathscr{G}_{\mu\nu} \equiv R^{(4)}_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R^{(4)} = 8\pi G_N T_{\mu\nu} = \frac{T_{\mu\nu}}{\overline{M}_{\rm Pl}^2} \tag{6.82}$$

where $\mathscr{G}_{\mu\nu}$ is the Einstein tensor. $T_{\mu\nu}$ on the right-hand side is the energy–momentum tensor made from the matter field and is the source of the gravity fields.

Going to the $D = 4 + \delta$ dimension, the metric is generalized to

$$ds^2 = G_{\rm MN} dx^M dx^N = \left(\eta_{\rm MN} + \frac{h_{\rm MN}}{2(\overline{M}_D)^{\frac{\delta}{2}+1}} \right) dx^M dx^N \tag{6.83}$$

where $\overline{M}_D$ is the Planck mass in the D-dimensional spacetime and is related with the 4D Planck mass by

$$\frac{1}{8\pi G_N} = \overline{M}_{\rm Pl}^2 = \overline{M}_D^2 (\overline{M}_D V)^\delta = \overline{M}_D^{2+\delta} (2\pi R)^\delta \tag{6.84}$$

where $V = 2\pi R$ is the volume in the ED. It is introduced to keep the mass dimension of the gravity field. The bulk action that gives the Einstein equation in D-dimension is given by

$$S_{\rm bulk} = \int d^{4+\delta}x \sqrt{-G} \left(\frac{\overline{M}_D^{2+\delta}}{2} R^{(4+\delta)} - \Lambda \right), \qquad G = \det G_{MN} \tag{6.85}$$

where the vacuum energy (aka cosmological constant) is separated for the sake of future discussions.

The Einstein equation in the higher dimension is expressed as [467]

$$\mathscr{G}_{AB} \equiv R^{(D)}_{AB} - \frac{1}{2} g_{AB} R^{(D)} = \frac{T_{AB}}{(\overline{M}_D)^{2+\delta}}, \qquad R^{(D)} = g^{AB} R^{(D)}_{AB} \tag{6.86}$$

Expressed in terms of h_{MN}

$$(\overline{M}_D)^{\frac{\delta}{2}+1}\mathscr{G}_{AB} = \partial_M\partial^M h_{AB} - \partial_A\partial^C h_{CB} - \partial_B\partial^C h_{CA} + \partial_A\partial_B h^C_C \\ - \eta_{AB}\partial_C\partial^C h^D_D + \eta_{AB}\partial^C\partial^D h_{CD} = \frac{T_{AB}}{(\overline{M}_D)^{\frac{\delta}{2}+1}} \tag{6.87}$$

For matters residing only in 4D, $T_{AB} = \eta^\mu_A\eta^\nu_B\, T_{\mu\nu}\prod_{j=1}^{\delta}\delta(y_j)$.

For convenience, we define the higher dimensional coordinates as

$$x^M \equiv (x\,;y) \equiv (x^\mu; y_j) \equiv (x^0, x^1, x^2, x^3; y_1, \dots y_\delta) \tag{6.88}$$

where the higher dimensions are compactified and y_j's are periodic $[f(y_j + 2\pi R) = f(y_j)]$. Expanding h_{AB} in Fourier series

$$h_{AB}(x;y) = \sum_{m_1=-\infty}^{+\infty}\cdots\sum_{m_\delta=-\infty}^{+\infty} h^{(m)}_{AB}(x)\frac{e^{i\sum\frac{m_j y_j}{R_j}}}{\sqrt{\prod_j 2\pi R_j}} \tag{6.89}$$

where $h^{(m)}$ is a shorthand for $h^{(m_1, m_2, \dots, m_\delta)}$. We separate the massive KK tower from the massless zero mode by

$$h^{(m)}_{AB}(x) = h^{(0)}_{AB} + h^{(\vec{k})}_{AB}, \qquad \vec{k} \equiv \left(\frac{j_1}{R_1}, \frac{j_2}{R_2}, \dots, \frac{j_\delta}{R_\delta}\right) \tag{6.90}$$

6.5.1
Decomposition of the Gravitational Fields

Let us consider how many components the D-dimensional gravity field has. A symmetric $D \times D$ tensor has formally $D(D+1)/2$ components. However, dynamics of the observables are not changed by coordinate translations, $x^M \to x^M + \xi^M$. That is, they have gauge freedom $h_{MN} \to h_{MN} + \partial_M\xi^N + \partial_N\xi^M$. We can impose D conditions to fix the gauge, for example, harmonic gauge, $\partial_M h^M_N = \frac{1}{2}\partial_N h^M_M$. There still remains extra residual gauge freedom with functions ξ^N, which satisfy $\partial_M\partial^M\xi^N = 0$ allowing another D conditions to be imposed. Therefore, the total number of degrees of freedom becomes

$$n_f \equiv \frac{1}{2}D(D+1) - 2D = \frac{1}{2}D(D-3) \tag{6.91}$$

This condition applies to each (m) mode and is independent of zero or finite mass.

For $D = 4$, we have $n_f = 2$, which reproduces the number of degrees of freedom for spin-2 massless field. Next, consider the case $D = 5$. Here, $n_f = 5$ and the metric can be expressed as

$$ds^2 = G_{MN}dX^M dX^N = g_{\mu\nu}dx^\mu dx^\nu + 2g_{\mu 5}dx^\mu dy + g_{55}dy^2 \tag{6.92}$$

and the gravitational fields are decomposed to $h_{MN} = h_{\mu\nu} \oplus h_{\mu 5} \oplus h_{55}$. For the $n = 0$ massless mode, the number of degrees of freedom is 2 for $h_{\mu\nu}$, 2 for $h_{\mu 5}$, and 1 for h_{55}. For the $n \neq 0$ KK tower, they are massive and the spin-2 $h^{(n)}_{\mu\nu}$ has five polarizations and hence $n \neq 0$ modes of $h^{(n)}_{\mu_5}$ and $h^{(n)}_{55}$ are eaten by $h^{(n)}_{\mu\nu}$.

Similarly, it is convenient to decompose a D-dimensional tensor $h^{(m)}_{MN}$ into a 4D tensor $h^{(m)}_{\mu\nu}$, 4D vectors $h^{(m)}_{\mu j}$, and 4D scalars $S^{(m)}_{ij}$, where $\mu, \nu = 0 \sim 3$, $i, j = 1 \sim \delta$. The decomposition can be expressed as

$$\left[h_{AB}\right] = \left[\begin{array}{c|ccc} h^{(m)}_{\mu\nu} & h^{(m)}_{\mu 5} & \cdot & h^{(m)}_{\mu D} \\ \hline h^{(m)}_{5\nu} & h^{(m)}_{55} & \cdot & h^{(m)}_{5D} \\ \cdot & \cdot & \cdot & \cdot \\ h^{(m)}_{D\nu} & h^{(m)}_{D5} & \cdot & h^{(m)}_{DD} \end{array}\right] = \left[\begin{array}{c|c} G^{(0)}_{\mu\nu} & V^{(0)}_{\mu j} \\ \hline V^{(0)}_{i\nu} & S^{(0)}_{ij} \end{array}\right] + \left[\begin{array}{c|c} G^{(\vec{k})}_{\mu\nu} & V^{(\vec{k})}_{\mu j} \\ \hline V^{(\vec{k})}_{i\nu} & S^{(\vec{k})}_{ij} \end{array}\right] \tag{6.93}$$

Generalization of the number count to $D = 4 + \delta$ is straightforward. For the $m = 0$ mode, there are one 4D graviton, δ massless vectors, and $\delta(\delta + 1)/2$ scalars to fill the total of $D(D - 3)/2 = (\delta + 1)(\delta + 4)/2$ degrees of freedom. For the $m \neq 0$ mode, there is one massive graviton with five polarizations, which eats one vector and one scalar. That leaves $\delta - 1$ massive vectors that eat $\delta - 1$ scalars. The total number of vector polarizations is $3 \times (\delta - 1)$. Therefore, the number of remaining massive scalars is $(\delta + 1)(\delta + 4)/2 - 5 - 3 \times (\delta - 1) = \delta(\delta - 1)/2$.

To derive field equations for the 4D tensors, vectors, and scalars of the KK tower separately, we modify gauge conditions slightly in a form convenient for the decomposed components. The number of degrees of freedom of a 4×4 symmetric tensor is 10. To decrease it to 5, which is the number of polarization states of the massive spin-2 tensor, two conditions are applied:

$$\partial^{\mu} G^{(\vec{k})}_{\mu\nu} = 0, \qquad G^{(\vec{k})\mu}_{\mu} = 0 \tag{6.94}$$

The increase from 2 (for the massless spin-2 tensor field) to 5 (massive tensor in KK tower) is achieved by eating one massless vector field and one massless scalar field. The remaining $\delta - 1$ massive vector fields that have $3 \times (\delta - 1)$ degrees of freedom had $4 \times \delta$ degrees of freedom formally. The reduction can be achieved by imposing the constraints

$$\vec{k}^{j} V^{(\vec{k})}_{\mu j} = 0, \qquad \partial^{\mu} V^{(\vec{k})}_{\mu j} = 0 \tag{6.95}$$

The second constraint is the conventional Lorentz condition. As the one massive spin-2 field and $\delta - 1$ massive vector fields have eaten δ scalars out of $\delta(\delta + 1)/2$ scalars, the following constraint is imposed on the scalars to reduce the number:

$$\vec{k}^{j} S^{(\vec{k})}_{jk} = 0 \tag{6.96}$$

Finally, out of $\delta(\delta-1)/2$ scalar fields, a special field referred to as the *radion* (denoted as $H^{(\vec{k})}$) is extracted from the trace part of the scalar field.

$$H^{(\vec{k})} \equiv \frac{1}{\kappa} S_j^{(\vec{k})j}, \qquad \kappa \equiv \sqrt{\frac{3(\delta-1)}{\delta+2}} \;^{19)} \tag{6.97}$$

The rest of the scalar fields are traceless ($S_j^{(\vec{k})j} = 0$). Applying the decomposition Eq. (6.93) to the Einstein equation Eq. (6.87) and using the auxiliary conditions discussed above, one can derive the field equations for each component [546, 547]:

$$(\partial_\mu \partial^\mu + k^2)\, G_{\mu\nu}^{(\vec{k})} = \frac{1}{\overline{M}_{\text{Pl}}}\left[T_{\mu\nu} - \left(\eta_{\mu\nu} + \frac{\partial_\mu \partial_\nu}{\vec{k}^2}\right)\frac{T_\lambda^\lambda}{3}\right] \tag{6.99a}$$

$$(\partial_\mu \partial^\mu + k^2)\, V_{\mu j}^{(\vec{k})} = 0 \tag{6.99b}$$

$$(\partial_\mu \partial^\mu + k^2)\, S_{ij}^{(\vec{k})} = 0 \tag{6.99c}$$

$$(\partial_\mu \partial^\mu + k^2)\, H^{(\vec{k})} = -\frac{\kappa}{2\overline{M}_{\text{Pl}}} T_\mu^\mu \tag{6.99d}$$

where

$$k^2 \equiv \sum_{j=1}^{\delta} \left|\frac{m_j}{R}\right|^2 \tag{6.100}$$

We can see why the radion is special. It defines the overall size of the compact dimension and is also the only field besides the 4D graviton that couples to the matter (energy momentum tensor) in 4D. The rest of the KK towers can be dropped from the following discussions.

6.6
Warped Extra Dimension

6.6.1
Anti-de Sitter Space AdS_5

So far, we thought that the ED was compactified because it is not observable. In consideration of the compactification size, we concluded that $\delta = 1$ and possibly $\delta = 2$ are excluded experimentally [see Eq. (6.2)]. There is a novel idea that the extra dimension need not be compactified and could even be infinite. One wonders how

---

19) The radion can be considered as the excitation of the radial field where the D-dimensional metric is decomposed as

$$G_{MN} dx^M dx^N = (\eta_{\mu\nu} + \tilde{h}_{\mu\nu}) dx^\mu dx^\nu - r^2 d\Omega_\delta^2 \tag{6.98}$$

In the definition of Eq. (6.97), the radion for $\delta = 1$ does not exist. However, by parameterizing y in the 5D as $y = r\theta,\ 0 \le \theta < 2\pi$, we still can talk of radial excitation. We will introduce the $\delta = 1$ radion in warped space in Eq. (6.113).

can it escape observations. The idea is to introduce a warped space known as the *Randall–Sundrum model* [521,522]. One starts with the 5D spacetime where the fifth space is filled only with vacuum energy. One introduces a metric parameterized as

$$ds_5^2 = G_{MN}dx^M dx^N = e^{-2\sigma(y)}g_{\mu\nu}dx^\mu dx^\nu - dy^2 \tag{6.101}$$

where $e^{-2\sigma(y)}$ is referred to as the *warp factor*. The form is motivated by the Robertson–Walker metric of the inflationary universe in 4D where the spacetime is filled with vacuum energy only (the so-called de Sitter space) [see Eq. (8.55)].

$$ds^2 = dt^2 - R(t)^2\eta_{\mu\nu}dx^\mu dx^\nu, \qquad R(t) \sim e^{\sqrt{\frac{\Lambda}{3}}t} \tag{6.102}$$

If the vacuum energy permeates in 5D but the matter does not, the metric would assume the same form with the fifth dimension y assuming the role of time in 4D. We will show that the metric Eq. (6.101) is indeed a solution to the Einstein equation.

The 5D action is given by

$$S_{5D} = \int d^4x \int_{-\pi R}^{+\pi R} dy \left[\sqrt{-G}\left(\frac{\overline{M}_5^3}{2}R^{(5)} - \Lambda\right)\right] + S_{\text{matter}} \tag{6.103}$$

where Λ is the cosmological constant term in 5D. Notice that the mass dimension of $R^{(5)}$ is the same as that of $R^{(4)}$. It is given as a polynomial of the second derivatives of the metric tensor. $\overline{M}_5$ is the Planck mass in 5D. We consider that matters are localized at two branes at $y = y_1, y_2$ $(y_1 = 0,\ y_2 = \pi R)$. The 4D metric $ds_{(i)}^2$ $(i = 1, 2)$ at each brane is defined by

$$ds_{(i)}^2 = g_{\mu\nu}^{(i)}dx^\mu dx^\nu = e^{-2\sigma(y_i)}g_{\mu\nu}dx^\mu dx^\nu \tag{6.104}$$

The action for the matter is given by

$$\begin{aligned} S_{\text{matter}} &= \sum_{i=1,2}\int d^4x dy\, \delta(y - y_i)\sqrt{-g^{(i)}}\left(\mathscr{L}^{(i)} - V^{(i)}\right) \\ g^{(i)} &= \det[g_{\mu\nu}^{(i)}] \end{aligned} \tag{6.105}$$

where the potential energy $V^{(i)}$ was separated, which acts as a gravitational source even in the absence of particle excitations. In the following, the details of the 4D Lagrangian $\mathscr{L}$ are irrelevant, so we set $\mathscr{L}^{(i)} = 0$.

Then, the equation of motion leads to the Einstein equation in vacuum.

$$R_{MN}^{(5)} - \frac{1}{2}G_{MN}R^{(5)} = \overline{M}_5^{-3}\left[G_{MN}\Lambda + \frac{\sqrt{-g}}{\sqrt{-G}}\delta_M^\mu\delta_N^\nu G_{\mu\nu}\sum_{i=1,2}V^{(i)}\delta(y - y_i)\right] \tag{6.106}$$

Plugging the metric tensor Eq. (6.101) into Eq. (6.106), one finds [521,548,549]

$$\sigma'^2 = \left(\frac{d\sigma}{dy}\right)^2 = -\frac{\Lambda}{6\overline{M}_5^3} \equiv k^2 \tag{6.107a}$$

$$\sigma'' = \frac{V^{(1)}}{3\overline{M}_5^3}\delta(y) + \frac{V^{(2)}}{3\overline{M}_5^3}\delta(y - \pi R) \tag{6.107b}$$

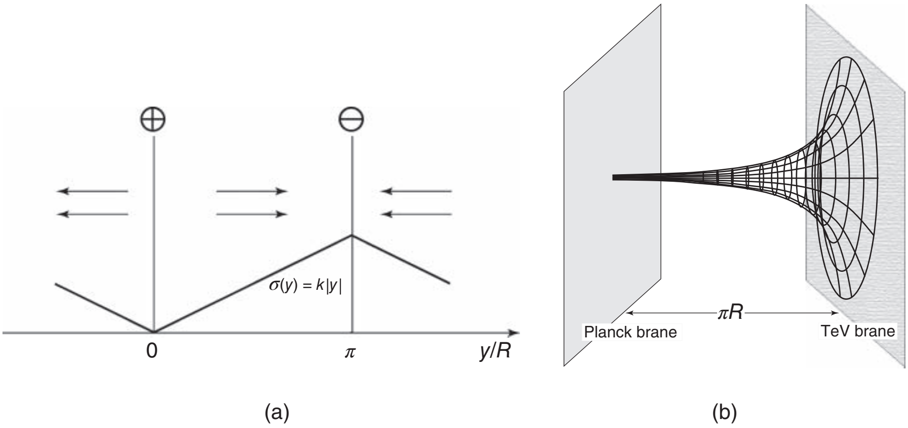


Figure 6.9 (a) Warp factor solution satisfying the orbifold configuration. Two branes at ⊕, ⊖ are referred as the *Planck and TeV brane*, respectively. (b) Warped space: rescaling across the fifth dimension. The gravitational field flux per unit area is diluted because of rapid expansion of the space. Gravity is strong at the Planck brane, but is weak at the TeV brane by a factor $e^{-k\pi R}$.

These are the equations for $\sigma(y)$. The first equation constrains the cosmological constant Λ to be negative, that is, the 5D bulk spacetime has to be a slice of the anti-de Sitter space (denoted as AdS_5). The function has the form $\sigma(y) = \sigma_0 + k|y|$ and is depicted in Figure 6.9a.

The additive constant σ_0 can be dropped by rescaling x^μ. Periodicity and continuity of the function on the orbifold require

$$V^{(1)} = -V^{(2)} = 6k\overline{M}_5^3 \tag{6.108}$$

Notice that, without the kinetic terms, the energy $V^{(1)}$, $V^{(2)}$ represents tension of the branes. Equation (6.108) means, if one brane has positive tension, the other has negative tension. Thus, the vacuum metric solution is given by

$$\begin{aligned} ds_5^2 &= e^{-2ky}\eta_{\mu\nu}dx^\mu dx^\nu - dy^2, \qquad 0 \le y \le \pi R \\ (ds_4^{(1)})^2 &= \eta_{\mu\nu}dx^\mu dx^\nu, \qquad (ds_4^{(2)})^2 = e^{-2k\pi R}\eta_{\mu\nu}dx^\mu dx^\nu \end{aligned} \tag{6.109}$$

The above metric is that of Minkowski spacetime (no cosmological constant term in 4D). To summarize, starting from the 5D bulk with the cosmological constant, and two 4D branes with opposite tensions, we have achieved the construction of a spacetime flat in 4D but warped in the fifth dimension. Warping is due to gravity in the fifth dimension created by the cosmological constant.

The warp factor means an expanding space toward the brane at $y = \pi R$ (Figure 6.9b). Analogy with the expanding universe helps to understand it intuitively. The metric ds_5^2 is reminiscent of the Robertson–Walker metric with dominant cosmological constant energy. There, the universe expands exponentially in time. Replacing time with the fifth coordinate y, one gets the exponentially expanding (or contracting depending on the sign of k) in y.

6.6.2
RS1 Scenario

We will show how the big hierarchy problem can be solved in the so-called RS1 scenario of the warped 5D spacetime. We consider a case $k > 0$, namely $V^{(1)} > 0$, $V^{(2)} < 0$, and assume that a 4D Higgs field (and other SM particles) is localized in the $y = \pi R$ brane. The brane at $y = 0$ is referred to as the *Planck (or UV: ultraviolet) brane* and that at $y = \pi R$ is referred to as the *TeV (or IR: infrared) brane*. The meaning of the name will become clear soon. Let us consider the action of a Higgs in the TeV brane.

$$\begin{aligned} S_{\text{Higgs}} &= \int d^4x \sqrt{-g^{(2)}} \left[g^{(2)\,\mu\nu} \partial_\mu H \partial_\nu H - \lambda \left(|H|^2 - v_0^2 \right)^2 \right] \\ g^{(2)}_{\mu\nu} &= e^{-2k\pi R} g^{(0)}_{\mu\nu}, \quad g^{(2)} = \det g^{(2)}_{\mu\nu} = (e^{-2k\pi R})^4 g^{(0)} = e^{-8k\pi R} g^{(0)} \end{aligned} \tag{6.110}$$

$g^{(2)}_{\mu\nu}$ represents the interaction with gravity. Expressing the 4D Higgs action in terms of $g^{(0)}_{\mu\nu}$, we have

$$S_{\text{Higgs}} = \int d^4x \sqrt{-g^{(0)}} \left[e^{-2k\pi R} g^{(0)\,\mu\nu} \partial_\mu H \partial_\nu H - e^{-4k\pi R} \lambda \left(|H|^2 - v_0^2 \right)^2 \right] \tag{6.111a}$$

where we have used $g^{(2)\mu\nu} = e^{2k\pi R} g^{(0)\,\mu\nu}$. We can renormalize the Higgs wave function such that $H \equiv \hat{H} e^{k\pi R}$ and obtain

$$S_{\text{Higgs}} = \int d^4x \sqrt{-g^{(0)}} \left[g^{(0)\,\mu\nu} \partial_\mu \hat{H} \partial_\nu \hat{H} - \lambda \left(|\hat{H}|^2 - \left(v_0 e^{-k\pi R} \right)^2 \right)^2 \right] \tag{6.111b}$$

A remarkable thing is that the bare VEV gets renormalized to the physical VEV by

$$v \equiv e^{-k\pi R} v_0 \tag{6.112}$$

If the Planck mass is the fundamental mass scale such that $v_0 \sim O(\overline{M}_{\text{Pl}})$, the physical VEV of the Higgs field is scaled down by the factor $e^{-k\pi R}$. A modest value of $kR \sim 12$ would reduce the Planck scale to the TeV scale. Note that the quartic coupling λ, which is expected of the order ~ 1, is unchanged. As the Higgs VEV determines the weak scale, the result is completely general. Any mass parameter on the TeV brane in the fundamental higher dimensional theory will be scaled down by the same factor. Thus, starting from a 5D theory with fundamental input parameters at the Planck scale, we have achieved the Planck/weak ($10^{19}/10^3$ GeV) hierarchy by merely imposing a modest $k\pi R \sim O(10)$ adjustment. There is no hierarchy problem between the Planck mass and the physical mass in 5D theory.

Radion So far we treated the radius in the fifth dimension as a fixed given constant. However, we are talking of relatively large value of $kR \sim O(10)$. The quantum excitation of the metric may not be negligible and we have to worry about the stability of the vacuum in 5D. The RS metric with quantum excitation is defined by

$$\begin{aligned} ds_5^2 &= e^{-2k|\phi|T(x)} g_{\mu\nu} dx^\mu dx^\nu - T^2(x) d\phi^2 \\ g_{\mu\nu} &= <g_{\mu\nu}> + (g_{\mu\nu} - <g_{\mu\nu}>) \equiv \eta_{\mu\nu} + \tilde{h}_{\mu\nu}, \\ T(x) &= <T> + (T(x) - <T>) = R + (T(x) - R) \end{aligned} \tag{6.113}$$

$<\cdots>$ denotes the vacuum expectation value and gravitons are excitations from the vacuum. Spacetime coordinates are no longer given environments but have become dynamical hence its dependence on x. T(x) is a scalar field referred to as the modulus field or conventionally as the radion.20) The vacuum expectation value of the modulus field defines the classical radius in the fifth coordinate. For the 4D part, $\tilde{h}_{\mu\nu} \ll 1$ and linearlization of the field is known to be an excellent approximation. Therefore, the back reaction which is of higher order in the gravitational field is negligible.

As it stands, the radion is massless because only derivative terms exist in the Einstein equation. It has no KK tower as they are eaten by the KK graviton (see statement after Eq. (6.92)). But the massless radion means additional contribution to the Newton potential (the fifth force) and is phenomenologically forbidden. It also means that the radion has no potential hence no stablizing power for the radion which is an oscillating radius. The radion has to have mass.

Stability of the radius in y is achieved by adding a massive scalar field in 5D bulk [546, 551]. It is a mechanism similar to the domain localized fermions that we discussed in Section 6.3.2. The interaction of the radion with the scalar field generates a potential. In simple words, it is a result of balancing between the kinetic term which wants to expand the bulk size and the mass term that wants to shrink it. As a result, the distance between the 4D branes is dynamically determined and stabilized. The potential generates mass of the radion comparable to that of KK gravitons. In the following we mainly discuss KK gravitions and mention to the radion when appropriate.

6.6.3
RS2 scenario

In the previous section, we worked out the hierarchy problem in the warped gravity background. In the following which we will refer to as the RS2 scenario, we turn on the KK excitation modes of the gravity and see their effect on the brane. The boundary condition of the RS2 is different from RS1. In RS1, a brane with positive tension resides at $y = 0$ (Planck brane) and the other, the SM world, with negative tension resides at $y = \pi R$ (IR brane). In RS2, we show that the IR brane can be taken to infinity [522, 548, 551]. That is, the main purpose of the RS2 scenario is

---

20) The canonically normalized radion field $r(x)$ is related with the modulus field $T(x)$ by [550]

$$\begin{aligned} r(x) &= \varphi(x) - <\varphi> = \Lambda_r (e^{-k\pi(T(x) - <T>)} - 1), \\ \Lambda_r &\equiv <\varphi> = \sqrt{\frac{6\overline{M}_5^3}{k}}\, e^{-k\pi R} \sim O(\text{TeV}) \end{aligned} \tag{6.114}$$

to show that the 5th dimension needs not be confined to a finite volume. This is in sharp contrast to the flat large extra dimension discussed in the beginning of this chapter. What we are going to show is that the zero mode of the graviton is localized at $y = 0$, but all the KK towers are pushed off to the IR brane and their net effect on the Planck brane is negligible enabling to take the IR brane to infinity. Whether we want to place it at infinity or set it at some fixed point of the orbifold is a choice of model building.

Let us see what effective strength appears in 4D under the given 5D metric as expressed by Eq. (6.109). First, we consider only the zero mode of the graviton that does not depend on y. The 5D action can be expressed as

$$\begin{aligned} S_5 &= \frac{\overline{M}_5^3}{2} \int d^4x \int_{-\pi R}^{+\pi R} dy \sqrt{-G}\, R^{(5)} \\ &= \overline{M}_5^3 \int d^4x \int_0^{\pi R} dy \sqrt{-g}\left[A R^{(4)} + \cdots\right] = S_{4\,\mathrm{eff}} + \cdots \end{aligned} \tag{6.115}$$

The 5D Ricci curvature $R^{(5)}$ is derived from the 5D metric and the $R^{(4)}$ from the 4D metric. The Ricci curvature is a combination of the second derivatives of the metric tensor. (See expressions in Appendix Eq. (E.17).)

$$\begin{aligned} R^{(5)} &= G^{MN} R_{MN} = G^{MN} R^K_{MKN} = G^{MN}\left[\partial_M \Gamma^K{}_{KN} + \cdots\right] \\ &= G^{MN}\left[\partial_M\left\{G^{KL}\left(\partial_L G_{KN} + \cdots\right) + \cdots\right\}\right] \end{aligned} \tag{6.116}$$

where R^J_{MKN} and Γ^J_{KN} are 5D Riemann curvature tensor and Christoffel connection in 5D. We need only to worry about the $MN = \mu\nu$ part to compare it with $R^{(4)}$. Terms containing $M,\ N = 5$ will contribute only to the potential after y integration. As coefficients of each term is identical for 4D and 5D curvatures, derivatives in Eq. (6.116) produce no extra factors (i.e., $\partial_\lambda G_{\mu\nu} = e^{-2ky}\partial_\lambda g_{\mu\nu}$). We can calculate A in Eq. (6.115) exactly by comparing the coefficient of any term with $MN = \mu\nu$. Using $G = (e^{-2ky})^4 g$, $G^{\mu\nu} = e^{2ky} g^{\mu\nu}$, $G_{\mu\nu} = e^{-2ky} g_{\mu\nu}$, and the first term in the second line of Eq. (6.116), we obtain

$$A = \sqrt{(e^{-2ky})^4}\,(e^{2ky})(e^{2ky})e^{-2ky} = e^{-2ky} \tag{6.117}$$

leading to

$$\begin{aligned} S_{4\,\mathrm{eff}} &= \overline{M}_5^3 \int d^4x \left(\int_0^{\pi R} dy\, e^{-2ky}\right)\sqrt{-g}R^{(4)} \\ &= \frac{\overline{M}_5^3}{2k}\left(1 - e^{-2k\pi R}\right)\int d^4x \sqrt{-g}R^{(4)} \end{aligned} \tag{6.118}$$

Comparing Eq. (6.118) with Eq. (6.81), we obtain

$$\begin{aligned} 16\pi G_N &= \frac{2}{\overline{M}_{\mathrm{Pl}}^2} = \frac{1}{\overline{M}_5^3}\frac{2k}{1 - e^{-2k\pi R}} \\ \overline{M}_{\mathrm{Pl}}^2 &= \frac{\overline{M}_5^3}{k}\left(1 - e^{-2k\pi R}\right) \end{aligned} \tag{6.119}$$

Contrast this with the relation $\overline{M}_{\rm Pl}^2 = \overline{M}_5^{2+\delta} V^\delta$ for the large, flat extra-dimensional relation [see Eq. (6.2)]. We see that there is a well-defined value for $\overline{M}_5$ in the warped extra space even in the $R \to \infty$ limit.

6.6.4
Gravitons in the RS Model

Arguments in the previous section used only the zero mode of the KK tower. We need to prove that the inclusion of all the higher KK towers does not affect the Newton's law of gravitation in 4D. In a warped space scenario, we can show that the KK gravitons are pushed off from the Planck brane localizing in the vicinity of the IR brane, and that they have masses $\sim ke^{-k\pi R}$. As $kR \sim O(10)$, the KK mass will be around ~TeV. In order to investigate the gravity in the bulk, we need to retain the $M,\ N = 5$ part in Eq. (6.116). It turns out that the field equation for the physical gravitational wave $h_{MN}(x, y)$ is the same for the massless 5D scalar field except for the normalization factor. [see Eq. (E.28)]. The 5D action for the scalar field can be expressed as

$$\begin{aligned} S_{5\,\rm scalar} &= \frac{1}{2}\int d^4x \int_{-\pi R}^{\pi R} dy\, \sqrt{-G}\, G^{MN} \partial_M \Phi \partial_N \Phi \\ &= \frac{1}{2}\int d^4x \int_{-\pi R}^{\pi R} dy\, e^{-4\sigma(y)} \sqrt{-g} \left[e^{2\sigma(y)} (\partial_\mu \Phi)(\partial^\mu \Phi) - (\partial_y \Phi)^2 \right] \end{aligned} \tag{6.120}$$

If we consider only the zero mode, the second term vanishes and the zero mode gravitational field reproduces Eq. (6.118) with $\overline{M}_5^3 R^{(4)}$ replaced with the scalar field Lagrangian. As $\partial_y \Phi$ is an odd function of y, compactification on the orbifold requires the boundary condition

$$\partial_y \Phi(x, y) = 0, \qquad \text{at} \quad y = 0,\ \pi R \tag{6.121}$$

This is nothing but the Neumann condition for the free-end string. Expanding Φ in terms of the KK tower,

$$\begin{aligned} &\Phi(x, y) = \sum_{n \neq 0} \psi_n(x) \varphi_n(y) \\ &(\partial_\mu \partial^\mu + m_n^2) \psi_n(x) = 0 \end{aligned} \tag{6.122}$$

To solve Eq. (6.120), we rewrite it in the following form:

$$S_{5\,\rm scalar} = \frac{1}{2}\int d^4x \int_{-\pi R}^{\pi R} dy\, e^{-2\sigma} \Phi \left[-M^2 - \partial_\mu \partial^\mu \right] \Phi \tag{6.123}$$

Equation (6.123) can be solved by using the complete system of the eigen equation

$$\begin{aligned} &M^2 \varphi_n(y) = m_n^2 \varphi_n(y) \\ &\quad M^2 \varphi_n = -e^{2\sigma} \partial_y \left[e^{-4\sigma} \partial_y \varphi_n(y) \right] = -e^{-2\sigma} \left[\partial_y^2 - 4\sigma' \partial_y \right] \varphi_n(y) \end{aligned} \tag{6.124}$$

It is necessary to normalize $\varphi_n(y)$ by

$$\int_{-\pi R}^{+\pi R} dy\, e^{-2\sigma} \varphi_n(y) \varphi_m(y) = \delta_{mn} \tag{6.125}$$

in order to obtain a correct kinetic energy term. Substituting $\varphi_n(y)$'s satisfying both (6.124) and (6.125), the right-hand side of Eq. (6.120) simplifies to

$$S_{5\,\text{scalar}} \to \frac{1}{2}\sum_n \int d^4x \left[\eta^{\mu\nu}\partial_\mu\psi_n\partial_\nu\psi_n - m_n^2\psi_n^2\right] \tag{6.126}$$

Setting $\varphi(y) = e^{2\sigma}\chi(y)$, Eq. (6.124) is rewritten as

$$\begin{aligned}&\left[-(\partial_y - 2\sigma')(\partial_y + 2\sigma')\right]\chi_n(y)\\ &= -\left[\partial_y^2 - 4\sigma'^2 + 2\sigma''\right]\chi_n(y) = m_n^2 e^{2\sigma}\chi_n(y)\end{aligned} \tag{6.127}$$

where $\sigma' = \text{sgn}(y)\,k$, $\sigma'' = 2k\big[\delta(y) - \delta(y - \pi R)\big]$. In order to grasp the qualitative features of the solution, we introduce new variables

$$z \equiv \text{sgn}(y)\frac{e^{k|y|} - 1}{k}, \qquad \hat{\chi}_n = \chi_n(y)e^{\frac{k|y|}{2}} \tag{6.128}$$

Then, using $e^{k|y|} = k|z| + 1$, $dz = (k|z| + 1)dy$, and so on, Eq. (6.127) reads [522]

$$\left\{-\frac{1}{2}\partial_z^2 + V(z)\right\}\hat{\chi}_n(z) = m_n^2\,\hat{\chi}_n(z) \tag{6.129a}$$

$$V(z) = \frac{15k^2}{8(k|z| + 1)^2} - \frac{3}{2}k\Big[\delta(z) - \delta(z - z_1)\Big] \tag{6.129b}$$

The shape of the potential is depicted in Figure 6.10. It is referred to as the *volcano potential.*

Solving Eq. (6.129) is a problem of quantum mechanics to determine energy eigenvalues. The delta function supports only one bound state localized at the origin, which is none other than our localized zero mode. As the potential is repulsive, all other states ($n \neq 0$ KK towers) are unbound and pushed toward $z = z_{\max} = z_1$. That is, the KK modes are more or less localized in the $y = \pi R$ brane. If the fifth dimension extends to infinity, the unbound KK modes will have continuous levels. On the orbifold, however, energy levels of the KK modes are

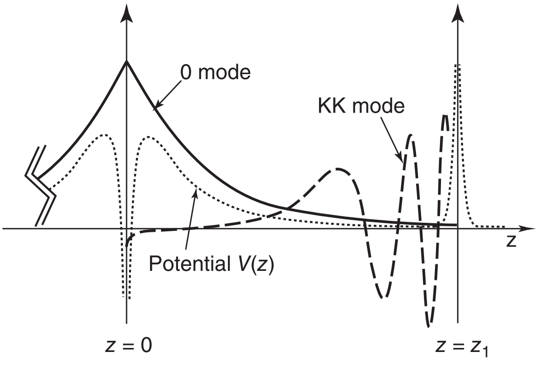


Figure 6.10 Graviton zero and Kaluza–Klein modes for the volcano potential (dotted line) on the orbifold. They are determined from an analogous one-dimensional quantum mechanics problem. Also shown are the localized zero mode (real line) and an example of additional KK mode wave function (dashed line).

quantized. The size $|z_{\max}| \sim e^{k\pi R}/k$ defines the level intervals and mass of the KK modes.

$$E_{\mathrm{KK}} \sim \frac{n}{z_{\max}} \quad \rightarrow \quad m_{KK}^2 \sim (nke^{-k\pi R})^2 \sim (n\overline{M}_{\mathrm{Pl}}e^{-k\pi R})^2 \sim O(\mathrm{TeV}^2) \tag{6.130}$$

That is, despite the fact that all the fundamental inputs in 5D setup are of the Planck scale (i.e., $k \sim \overline{M}_{\mathrm{Pl}}$), the KK mass will be of the TeV scale.

One may worry that the gravitational strength of matter may considerably be modified by interacting with an infinite number of KK gravitons. Each of the contribution, however, is of the Yukawa potential type because of the finite mass, and their sum would be finite using the same logic that leads to Eq. (6.15). The argument is valid if the brane at $z = z_1$ is pushed to an infinite distance. As the 4D gravitational constant in Eq. (6.119) is kept finite in this limit, it is possible to work in 5D framework with the fifth dimension infinitely extended. This is a special feature of the warped space scenario in contrast to the flat ED.

The arguments here have been qualitative, but one can make it more quantitative by explicitly solving the equations for the KK tower, which are known to be of Bessel function type.

Problem 6.1 *1) Show that by introducing $u = (m_n/k)e^{\sigma} = m_n(|z| + 1/k)$, Eq. (6.127) becomes [549]*

$$\left[\frac{\partial^2}{\partial u^2} + \frac{1}{u}\frac{\partial}{\partial u} + \left\{1 - \left(\frac{2}{u}\right)^2\right\}\right]\chi_n(u) = 0 \tag{6.131}$$

which is an equation for the Bessel function of the second order. Two types of Bessel functions behave like

$$\begin{aligned} &J_2(u) \sim \frac{u^2}{8}, \qquad Y_2(u) \sim -\frac{4}{\pi u^2} \qquad \text{for} \quad u \ll 1 \\ &J_2(u) + iY_2(u) \sim \sqrt{\frac{2}{\pi u}}\exp\left[i\left(u - \frac{5}{4}\pi\right)\right] \qquad \text{for} \quad u \gg 1 \end{aligned} \tag{6.132}$$

2) Show that the wave function is given by

$$\chi_n(u) \sim N_m\sqrt{u}\left[Y_2(u) + \frac{4k^2}{\pi m_n^2}J_2(u)\right], \qquad N_m \sim \frac{\pi m_n^{\frac{3}{2}}}{4k^2} \tag{6.133}$$

3) Show that the boundary condition $\partial_y\varphi_n(y) = 0$ translates to

$$(u\partial_u + 2)K_2(u) = zK_1(u) = 0 \qquad K_2 = J_2,\ Y_2 \tag{6.134}$$

Therefore, mass of the KK modes can be determined from the zeros of the Bessel function.

$$m_j = x_j\, ke^{-k\pi R}, \qquad J_1(x_j) = 0 \tag{6.135}$$

Approximately,

$$x_j \simeq \begin{cases} 3.8 & j = 1 \\ 7.0 & j = 2 \\ 10.2 & j = 3 \\ 16.6 & j = 4 \\ \vdots & \end{cases} \tag{6.136}$$

4) Show that the zero mode is not a Bessel function but can be obtained as $m \to 0$ *or equivalently* $u \to 0$ *limit.*

$$\chi_0(y) = \frac{1}{k(k|z|+1)^{\frac{3}{2}}} \sim \lim_{m\to 0} uY_2(u) \tag{6.137}$$

In summary, by adopting the warped fifth dimension and placing gravity in the Planck brane, we have achieved the TeV scale particle spectra at the TeV brane. There is no hierarchy problem here. This was RS1 scenario. In RS2, we have seen that localization of the 4D gravity (i.e., zero mode) at the Planck brane and the KK mode at the TeV brane was automatic with $m_{\text{KK}} \sim O(\text{TeV})$. What had to be assumed in many models to construct a viable phenomenology for LED is automatic in the warped space.

6.6.5
Signals for Warped ED

In the original RS1 scenario, the entire SM particles, that is, including fermions and gauge fields, are localized in the TeV brane. It was subsequently realized that, in order to solve the Planck-weak hierarchy, only the SM Higgs boson needs to be localized on or near the TeV brane. The SM fermions and gauge fields need not be confined in the TeV brane, which means that they can propagate in the bulk. We no longer need to restrict ourselves to assume that only the graviton can propagate in the bulk. Although the fundamental strength of the warped space in 5D is $\overline{M}_5^3/k \sim \overline{M}_{\text{Pl}}^2$ [Eq. (6.119)], both the SM particles and the KK gravitons turned out to have masses of O (TeV) [Eq. (6.130)]. While the zero-mode graviton is localized on the Planck brane, all the massive KK modes are repelled away from it and are localized near or on the TeV brane. The separation away from the Planck brane is critical. The warp factor has reduced the mass scale of the KK graviton as well as the SM particles to O(TeV).

Curiously, the same mechanism works to enhance the gravitational coupling strength, which becomes of the order TeV^{-1}. This is most easily recognized by looking at the normalized wave function of the KK mode $\varphi_n(y) = e^{\sigma(y)}\chi(y)$ see Eq. (6.125) (6.127) where $\chi_n(y)$ is expressed by the second-order Bessel functions.

This means that

$$\frac{\varphi(y)|_{y=\pi R}}{\varphi(y)|_{y=0}} \sim e^{k\pi R} \tag{6.138}$$

That is, the normalization constant is exponentially enhanced by the (inverse) warp factor. Hence, the full Lagrangian for the graviton excitations interacting with matter on the TeV brane is [550, 552]

$$\begin{aligned} \mathscr{L}_{\text{TeV}} &= -\frac{1}{\overline{M}_{\text{Pl}}} T^{\mu\nu} h^{(0)}_{\mu\nu} - \frac{1}{\Lambda_\pi} T^{\mu\nu} \sum_{n=1}^{\infty} h^{(n)}_{\mu\nu} + \frac{1}{\Lambda_r} T^{\mu}_{\mu} r \\ \Lambda_\pi &= \overline{M}_{\text{Pl}} e^{-k\pi R} \quad \Lambda_r = \sqrt{6}\Lambda_\pi \end{aligned} \tag{6.139}$$

where we added the radion interaction for comparison. This means that the coupling strength of the KK gravitons and the radion at the TeV brane is not the fundamental strength $1/\overline{M}^2_{\text{Pl}}$ but $1/\Lambda^2_\pi \sim 1/O(TeV)^2$. This is in sharp contrast to the flat ED in which the gravitational strength is universal to both the zero mode and the massive KK mode (see Eq. (6.172)). Consequently, the phenomenology of the warped space scenario is quite distinct from that of the flat ED.

The decay width of the graviton with mass m_G can be estimated on the grounds of dimensional arguments.

$$\Gamma_G \sim n_{\text{SM}} \frac{m_G^3}{\Lambda_\pi^2}, \qquad m_G = x_n k e^{-k\pi R} = x_n \Lambda_\pi \frac{k}{\overline{M}_{\text{Pl}}} \tag{6.140}$$

where $x_n s$ are zeros of the Bessel function given in Eq. (6.136) and n_{SM} is the number of SM particles into which the graviton can decay. This means that the KK gravitons in the warped space behave like new spin-2 resonances that can be produced and observed individually.[21] In particular, they will decay on a time scale of 1/TeV. Its detection is analogous to that of a new neutral gauge boson Z' discussed in the GUTs in Chapter 3. One can think of the KK gravitons appearing as resonances of pairs of SM particles in the reaction

$$q\bar{q},\, gg \to G^{(1)} + X \to \gamma\gamma,\, l\bar{l},\, q\bar{q},\, gg + X \tag{6.141}$$

Notice that, unlike Z, the graviton can couple to 2γs and $2g$s because of its democratic nature. Furthermore, the KK states are not evenly spaced.

If the first graviton KK state is observed, then parameters of this model can be uniquely determined from the location and width of the resonance. If the mass is given, the cross section can be determined as a function of the collider energy $\sqrt{s}$. A calculation of production cross section by [552, 553] is given in Figure 6.11.

A collider reach on m_1 (the first KK graviton) and k is given in Figure 6.12.

---

[21] The radion is most likely lighter than the KK gravitons in which case the radion would be detected first. The radion mixes with the Higgs after electroweak symmetry breaking, and its phenomenology would be very similar to that of the SM Higgs with the coupling strength reduced mostly by v(=248 GeV)/Λ_r. Notice, however, that the coupling to the gluon is much enhanced.

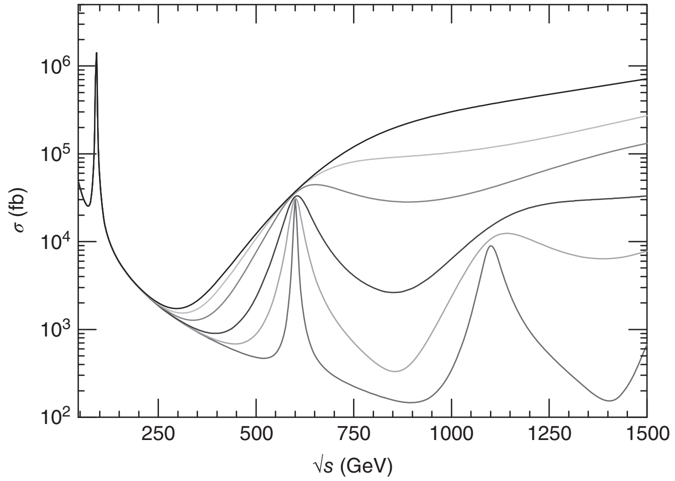


Figure 6.11 Cross section for $e^-e^+ \rightarrow \mu^-\mu^+$ including the exchange of a tower of KK gravitons, taking the mass of the first mode to be 600 GeV, as a function of $\sqrt{s}$. From top to bottom, the curves correspond to $k/\overline{M}_{\text{Pl}}$ = 1.0, 0.7, 0.5, 0.3, 0.2, and 0.1 (Reproduced with permission of [552–554].)

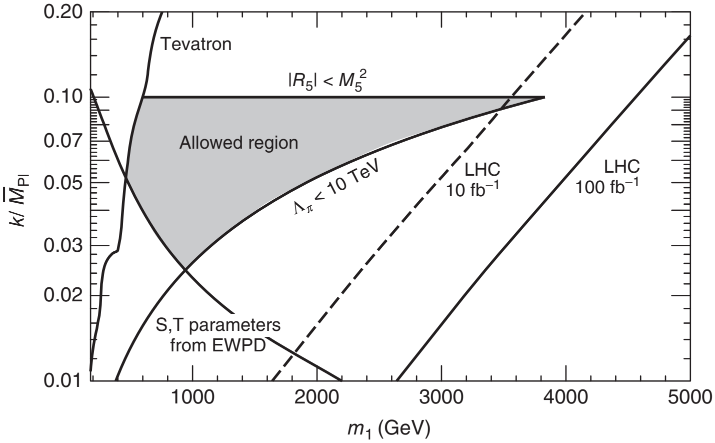


Figure 6.12 Summary of experimental and theoretical constraints on the RS model for the case where the SM lies on the TeV brane, in $k/\overline{M}_{\text{Pl}}$, and in m_1 (mass of the first excited KK graviton) plane. $R_5 = 20k^2$ is the Ricci scalar curvature (denoted as $R^{(5)}$ in the text) and $\Lambda_\pi = \overline{M}_{\text{Pl}} e^{-k\pi R}$. The allowed region lies in the center as indicated. The LHC sensitivity to graviton resonances in the Drell–Yan channel is represented by the diagonal dashed and solid curves, corresponding to 10 and 100 fb^{-1} of integrated luminosity, respectively. The theoretically allowed region will be drastically changed once a model of flavor such as the one in Figure 6.6 is included. (Reproduced with permission of [553, 555].)

6.7
Universal Extra Dimension (UED)

6.7.1
General Features

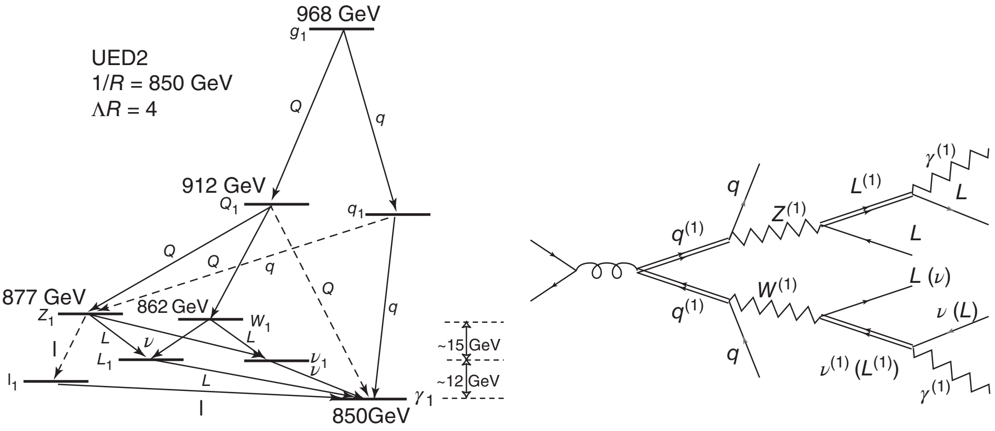
In the UED models [547, 556], all the matter fields as well as the gauge fields propagate in the bulk. The bulk is assumed to be flat as in the ADD model. An interesting feature of the model that results from the assumption is the KK parity (defined by $(-1)^n$) conservation rule. As the ED of the SM fields is at least of O(TeV) size, the KK excitation energy is much bigger than typical mass difference among the SM particles, making all the KK levels almost degenerate at the tree level. The KK parity conservation forbids mixing of the SM particles with the first level KK excitations; hence, radiative corrections to any electroweak precision observables are suppressed. The fact provides a natural explanation for the small hierarchy. Therefore, experimental constraints on UED are very weak and make the UED models phenomenologically very attractive.

Another interesting consequence of the KK parity conservation is that KK excitations are produced in pairs for n=odd levels though a single n=even level can also be excited. and that the lightest KK particle (LKP) is stable. That is, it plays a role quite similar to the R-parity conservation in supersymmetry. The LKP is weakly interacting and thus is an ideal candidate for the dark matter. As the energy gap of the KK excitation is large, production of KK particles will be limited almost exclusively to the first excited states. The produced KK particles will decay directly or in cascade to the LKP, emitting the SM particles (see Figure 6.13).

All the above considerations make the UED phenomenology quite similar to that of supersymmetric theory and make it hard to differentiate UED KK excitations from the SUSY particles. Unlike the SUSY, the KK excitations consist of the same spin partners of the SM particle. Therefore, one has to make spin analyses to differentiate the UED KK excitations from sparticles.

6.7.2
Selection Rules

Let us ignore gravity for a while. Then, interactions of the UED particles are quite similar to those of the SM particles. For simplicity, we consider a Lagrangian in 5D space. Extension to larger dimensions will be explained when necessary. The fifth dimension is an orbifold S^1/Z_2 to have chiral fermions. We denote a fermion with quantum number of ψ_L in the SM as ψ and that of ψ_R as ψ'. They have symmetry $\psi(x, y) = i\Gamma^5\psi(x, -y)$, $\psi'(x, y) = -i\Gamma^5\psi'(x, -y)$ to have the correct zero modes at the $y = 0$ brane. Denoting the gauge and Higgs field in 5D as B_M $(M = 1 \sim 5)$ and



(a) (b)

Figure 6.13 (a) Qualitative sketch of UED spectroscopy in the 4 + 1 extra dimensions [557]. The numbers indicate the mass values of a benchmark model. They are the level-1 KK spectroscopy depicting the dominant (solid lines), rare (dashed lines) transitions, and the resulting decay products. g_1, Q_1, q_1, $W_1^+/Z_1/\gamma_1\cdots$ are the first KK excitations of gluons, heavy quarks, light quarks, $W^+/Z/\gamma,\cdots$ [556, 558, 559] (b) Feynman diagram of a typical pair production of the first KK excitations of light quarks subsequently decaying to KK excitations of Z, W, ν, with the end product of LKP (denoted as $\gamma^{(1)}$). In the figure, L denotes either charged leptons (l) or neutrinos(ν_l) [557, 560]

Φ, the UED Lagrangian without gravity is given by

$$\begin{aligned}\mathscr{L} = &-\frac{1}{4}\sum_{\text{gauge}} F_{MN}F^{MN} + \sum_{\text{fermion}} i\overline{\psi}\Gamma^M D_M\psi + (D^M\Phi)^\dagger(D_M\Phi)\\ &+\sum_{\psi,\psi'}\lambda_5\overline{\psi}\Phi\psi' - V(\Phi)\end{aligned} \tag{6.142}$$

$$F_{MN} = \partial_M B_N - \partial_N B_M \tag{6.143a}$$

$$\psi = \frac{1}{\sqrt{\pi R}}\left[\psi_L^{(0)} + \sqrt{2}\sum_{n=1}^{\infty}\left\{\psi_L^{(+n)}\cos\frac{ny}{R} + \psi_R^{(-n)}\sin\frac{ny}{R}\right\}\right] \tag{6.143b}$$

$$\psi' = \frac{1}{\sqrt{\pi R}}\left[\psi_R^{'(0)} + \sqrt{2}\sum_{n=1}^{\infty}\left\{\psi_R^{'(+n)}\cos\frac{ny}{R} + \psi_L^{'(-n)}\sin\frac{ny}{R}\right\}\right] \tag{6.143c}$$

where $(\psi_{L,R}^{\pm(n)}, \psi_{L,R}^{'\pm(n)}) = P_{L,R}(\psi^{(\pm n)}, \psi^{'(\pm n)})$ and D_M is the covariant derivative. At the zero-mode level, $\psi^{SM} = \psi_L^{(0)} + \psi_R^{'(0)}$. At each n level, one gets two KK excitations $\psi^{(n)} = \psi_L^{(+n)} + \psi_R^{(-n)}$ and $\psi^{'(n)} = \psi_R^{'(+n)} - \psi_L^{'(-n)}$ with KK mass n/R. Note that the minus sign in the definition of $\psi^{'(n)}$ is chosen so that the mass term has the proper sign. $\psi^{(n)}$ and $\psi^{'(n)}$ correspond to KK excitations of ψ_L and ψ_R in the SM.

Let us see the interaction Lagrangian for the production of KK excitations, which is a product of a fermion pair and the gauge field B. To obtain the 4D interaction, one has to integrate over y.

$$\begin{aligned}\mathscr{L}_{\rm int} &= g_5 \int_0^{\pi R} dy \left[\overline{\psi}(x,y)\gamma^M \psi(x,y) B_M(x,y)\right]\\ &= g_5 \sum_{l,m,n\geq 0} \overline{\psi}^{(l)}\gamma^M \psi^{(m)} B_M^{(n)} \times \int_0^{\pi R} dy \frac{(\sqrt{2})^{\theta(l)+\theta(m)+\theta(n)}}{(\pi R)^{\frac{3}{2}}} \cos\frac{ly}{R}\cos\frac{my}{R}\cos\frac{ny}{R}\end{aligned} \tag{6.144}$$

where $\theta(i) = 0$ for $i = 0$ and 1 otherwise. The condition that the integral over y does not vanish imposes a condition which we refer to as the KK conservation rule.

$$|l \pm m \pm n| = 0 \tag{6.145}$$

The constraint is a result of momentum conservation but it is broken by S^1/Z_2 orbifold. If the rule holds at the tree level as is the case for minimum version of UED, it is broken by the radiative corrections because they generally induce two identical terms localized at fixed points ($y = 0, \pi R$), that is, terms proportional to $\delta(y) + \delta(y - \pi R)$. Still, a residual symmetry ($y \to \pi R - y$) exists between the two fixed points. Its application to Eqs. (6.143) together with the Z_2 symmetry make the KK parity $(-1)^n$ a conserved number [558]. Therefore, $n = 1$ KK levels have to be produced pair-wise but a single $n = 2$ KK level production is allowed. It also means that LKP, having nonzero KK number, is forbidden to decay to SM particles, that is, it is stable. The effect of the KK parity is quite similar to that of R-parrity in the supersymmetry. The 4D Lagrangian for the SM gauge particle to produce a pair of KK fermions becomes

$$\mathscr{L}_{\rm int}^{\rm 4D} = \frac{g_5}{\sqrt{\pi R}} \sum_m \int d^4x \overline{\psi}^{(m)} \left(\gamma^\mu B_\mu^{(0)} + i\gamma^5 B_5^{(0)}\right) \psi^{(m)} \tag{6.146}$$

One sees that this is an interaction Lagrangian for a gauge and Yukawa interaction with the 4D coupling constant $g_4 = g_5/\sqrt{\pi R}$. The fifth component of the gauge field B_5 appears as a scalar in 4D. As B_5 has odd Z_2 parity, $B_5^{(0)}$ is projected out under normal circumstance.

The masses of the KK excitations get two contributions: one is n/R, which comes from the ED, and the other from the Higgs VEV. However, as can be seen from Eq. (6.142), the Higgs couples the LH fermion ψ to the RH fermion ψ' and the resultant mass matrix is, in general, nondiagonal. The masses of the KK excitations, therefore, are obtained by diagonalizing the mass matrices

$$\mathscr{L}_{\rm mass} = \left[(\overline{\psi}^{(n)}, \overline{\psi}'^{(n)})\right] \begin{bmatrix} \frac{n}{R} & m_{\rm SM} \\ m_{\rm SM} & \frac{n}{R} \end{bmatrix} \begin{bmatrix} \psi^{(n)} \\ \psi'^{(n)} \end{bmatrix} \tag{6.147}$$

where $m_{\rm SM} = \lambda_5 <\Phi_0>$ is the mass of the SM fermions. The mass eigenstates for the fermion KK excitations will then contain an admixture of both the RH and LH fields. As the SM fermion masses are generally much smaller than O(TeV), we may neglect them and assume the KK excitations have mass given by $m_n = n/R$, in which case, all the KK excitations are degenerate. Of course, the degeneracy can be broken by radiative corrections. They are typically a few percent for the electroweak corrections and up to 30% for the QCD corrections [556].

6.7.3
Constraints

A typical problem of adding extra gauge bosons that couple to light fermions is that they give large contributions to electroweak observables such as m_Z, Γ_Z, G_F, and $\alpha_{em}(m_Z)$, which are severely constrained by precision data. The effect of additional one-loop radiative corrections generated by the presence of the ED particles can be parameterized by the Peskin–Takeuchi S, T variables (see Section 5.7 of [2]). S is an isospin-conserving parameter and T is an isospin-breaking one. According to calculations, the contributions to the T parameters are given by [523]

$$T = \sum_n D_n(T_n^V + T_n^h + T_n^t) \tag{6.148}$$

where suffixes V, h, t denote contributions from gauge bosons, Higgs, and (t,b) doublet. The sum is over all KK modes up to the cut-off scale Λ of the extra δ-dimensional theory, and D_n is the density of states. The individual contributions are

$$\begin{aligned}\alpha T_n^V &= -\frac{\alpha}{4\pi\cos^2\theta_W}\frac{(2\delta+1)m_W^2}{6m_n^2}\\ \alpha T_n^h &= -\frac{\alpha}{4\pi\cos^2\theta_W}\frac{5m_h^2+7m_W^2}{12m_n^2}\\ \alpha T_n^V &\simeq \frac{m_t^4}{8\pi^2 v^2 m_n^2}\end{aligned} \tag{6.149}$$

where $m_n = n/R$ is the KK mass. Using the experimental values for the SM parameters, the T parameter is roughly

$$T \simeq 0.76 \sum_n D_n \frac{m_t^2}{m_n^2} \tag{6.150}$$

A similar calculation for the S parameter gives

$$S \simeq 0.01 \sum_n D_n \frac{m_t^2}{m_n^2} \tag{6.151}$$

The KK contribution to T is two orders magnitude larger than that to S. Note that the sum over states is convergent for $\delta = 1$, but log divergent for $\delta = 2$ and power divergent for $\delta > 2$. Therefore, the UED models with $\delta > 2$ appear somewhat problematic. Closing our eyes to such problems, comparison of Eq. (6.150) with the electroweak precision data can give a lower bound on the inverse radius of the ED. A loose constraint $T \lesssim 0.4$ leads to $1/R \gtrsim 300\,(500)$ GeV for $\delta = 1\,(2)$. These bounds are probably underestimated but still much lower compared to non-UED models. The latter typically gives several TeV range [561].

Recent analysis taking into account the newly discovered Higgs gives the lower limit of 700 GeV [562–564]. WMAP (Wilkinson Microwave Anisotropy Probe) data on the abundance of the relic dark matter constrains $1/R$ from above [564, 565]. Therefore, there is a severe constraint on $1/R$ on the minimal UED.

$$700\ \text{GeV} \lesssim 1/R \lesssim 1600\ \text{GeV} \tag{6.152}$$

In the minimal UED, there are only three parameters, namely $1/R$, Λ, and m_{Higgs}. The cut-off parameter Λ is limited to $\Lambda R \lesssim 60$ from a naive dimensional argument, but is more severely constrained by the light Higgs mass observed at LHC and the vacuum stability to $\Lambda R \lesssim 6$ [566].

6.7.4
Signals for UED

UED Without Gravity From the discussions in the previous section, one sees that the main production mechanism in the colliders for obtaining the KK particles in UED will be given by pair production. As the coupling of the KK matter is the same as those of their SM partners, the most copiously produced modes in the hadron collider are the first-level excitations of gluons and quarks. Figure 6.13 shows a typical example of the first excited KK levels in UED and their decay patterns radiating the SM particles.

As LKP is stable and weakly interacting, it is the end product of decay chains and is observed as missing energy (ME_T) just like the neutrino or the lightest supersymmetric particle. Then, the first-level KK excitations of quarks and gluons produced at a hadron collider will undergo chain decays to the LKP, radiating SM particles in the process. Produced $W^{(1)}$, for instance, will decay via $W^{(1)} \to l\nu^{(1)} \to l + \nu$+LKP, which behaves just like a heavy W' boson that appeared in the discussion of the GUTs. (See Figure 3.7) As both the produced KK excitations and the LKP are relatively heavy, they will be produced with small Q values, as illustrated in Figure 6.13a for a benchmark model. A typical signal for the production of the KK pair would be several relatively soft leptons (or jets) plus large missing energy. ‘‘Relatively soft’’ means that their energy will be the difference $Q_q - W_1/Z_1$ or $W_1/Z_1 - \gamma_1$. If the mass difference is sufficiently small, the KK particle can be quasi-stable and may be observed as a long-living particle.

A typical pair production process is given by

$$pp \to q_1 q_1 + X \to ME_{\text{T}} + \text{jets} \tag{6.153a}$$

$$pp \to Q_1 Q_1 + X \to V_1 V_1' + \text{jets} \to 2 \sim 4\, l + ME_{\text{T}} \tag{6.153b}$$

where ME_{T} denotes the missing transverse energy. A typical production and the subsequent decay chain are shown in Figure 6.13-(b). Table 6.1 lists generic branches and their ratios of the first KK excitations in UED models.

As mentioned earlier, all the above signatures are very similar to those of supersymmetric particle productions. Consider, for instance, Z_1, W_1, and γ_1 as higher neutralino/chargino and LSP; then, one exactly has the experimental signature of the supersymmetry. To differentiate UED from SUSY, one needs to measure the spin of the particles. One naturally measures the angular distribution of emitted leptons, but it is not so easy at the hadron collider [567–569]. A clean signal may be obtained at the electron collider [570, 571].

UED with Gravity If the gravitational interaction is included, it breaks the KK number conservation and expands the class of possible signatures. However,

Table 6.1 First-level KK-state decay branches and branching ratios [556, 558, 559].

$B(g_1 \to Q_1 Q)$	$\sim 50\%$
$B(g_1 \to q_1 q)$	$\sim 50\%$
$B(q_1 \to q\gamma_1)$	$\sim 100\%$
$B(t_1 \to W_1 b, H_1^+ b)$	$\sim 100\%$
$B(Q_1 \to QZ_1 : W_1 : \gamma_1)$	$\sim 33\% : 65\% : 2\%$
$B(W_1 \to \nu L_1 : \nu_1 L)$	$1/6 : 1/6$ (per flavor)
$B(Z_1 \to \nu\nu_1 : LL_1)$	$1/6 : 1/6$ (per flavor)
$B(L_1 \to \gamma_1 L)$	$\sim 100\%$
$B(\nu_1 \to \gamma_1 \nu)$	$\sim 100\%$
$B(H_1^\pm \to \gamma\gamma_1,\ H^\pm \gamma_1)$	$\sim 100\%$

constraints on the ED by the gravity are much weaker than in the SM. If the size of the ED is as large as eV^{-1}, residence of the SM particles has to be constrained to regions smaller than TeV^{-1}, that is, on fat branes in the bulk [528].

As the gravity coupling is $1/\overline{M}_D \sim 1/\text{TeV}$, it is not much weaker than the SM particles. If the graviton interaction is included, the KK excitations produced at colliders can decay directly to their SM partners by radiating a graviton. The experimental signal would be 2 – jets with a large missing p_T (as they come from the decay of a heavy particle) plus missing energy (taken away by the gravitons) [560, 561, 572]. An even more interesting scenario can appear in the case when decays due to mass splitting as well as gravity interactions take place. Then, one could have, for example, excitations of a quark decaying to γ_1(=LKP) through the usual electroweak interaction followed by the γ_1 decay to a photon and a graviton through gravity-mediated interaction. The experimental signal for this case will be quite striking at the hadron collider, consisting of two large p_T photons plus the missing energy [573].

As the coupling of matter to gravitons does not obey KK number conservation rules, it is possible to produce single KK excitations of quarks or gluons at colliders. If the gravitons are exchanged virtually to pair-produce KK modes, signals will be 2-jets plus missing energy (or a jet plus photon/lepton with missing energy if the KK excitation decays first to the LKP). If the gravitons are emitted directly, the signal would be a single jet/photon/lepton plus missing energy (again depending on the decay pattern of the KK particles produced).

6.8
Searches for Generic ED

We now discuss experimental searches for generic ED models. Models based on the extra-dimensional theory in their original form may be classified into three big categories:

1) The original ADD-type scenario in which the space of the ED is large, flat, and compactified and only the gravitons reside in the bulk.
2) The UED where all the SM particles and the graviton reside in the bulk.
3) The warped space scenario by RS.

All of them have their own pros and cons, and a variety of models have been proposed as their mixtures. From phenomenological point of view, it is more convenient to categorize into (1) ADD-type large extra-dimensional models (referred to as *ADD*) and TeV^{-1} size ED, which includes RS, UED and part of ADD models. In the former, the level splitting can go down to eV^{-1} scale and the KK levels could be considered continuous from a practical point of view. In the latter, the KK level intervals are of $O(\text{TeV})$ scale and could show up as TeV-size resonances of the SM. Simple phenomenology of the RS and UED models was already described in each section. We describe here astrophysical/cosmological constraints and phenomenological formulas relevant for observing the ADD-type extra dimensions in the collider experiments. Experimental results using these formulas will be discussed afterwards.

6.8.1
Astrophysical Constraints on ADD Models

The mass splitting of light KK modes in LED is

$$\begin{aligned}\Delta m &\sim \frac{1}{R} \overset{(6.2\text{b})}{\sim} \overline{M}_D \left(\frac{\overline{M}_D}{\overline{M}_{\text{Pl}}}\right)^{\frac{2}{\delta}} = \left(\frac{\overline{M}_D}{\text{TeV}}\right)^{\frac{(\delta+2)}{\delta}} 10^{\frac{12\delta-31}{\delta}}\ \text{eV} \\ &\sim 10^{-4}\text{eV}\,(\delta=2) - 10\,\text{MeV}\,(\delta=6)\end{aligned} \tag{6.154}$$

It is extremely small in particle physics scale, and the KK spectrum can be treated as a continuum for any practical purpose.

If the mass of the KK gravitons are small, astrophysical objects with temperature $T > m_{\text{KK}}$ can emit them, which will reveal themselves as deviations from the standard explanations [547]. We can estimate the order of thermal graviton production rate (per unit time and volume) by multiplying the coupling strength of each graviton ($\sim 1/\overline{M}_{\text{Pl}}^2$) by the number of accessible modes $\left[\sim (T/\text{number of level intervals})^\delta \sim (TR)^\delta = T^\delta \overline{M}_{\text{Pl}}^2/\overline{M}_D^{\delta+2}\right]$.

$$\frac{dN}{dt} \propto \frac{1}{\overline{M}_{\text{Pl}}^2}(TR)^\delta \sim \frac{T^\delta}{\overline{M}_D^{\delta+2}} \tag{6.155}$$

One sees that the coupling of each graviton is small ($\sim 1/\overline{M}_{\text{Pl}}^2$), but the number of available states is large ($\sim \overline{M}_{\text{Pl}}^2$), which compensates the weakness of the coupling. Equation (6.155) shows that, if the available energy is comparable to the fundamental scale of the gravity, one gets observable effects in direct production or virtual exchange processes. It also shows that the effect is larger as the temperature goes higher, so one seeks high-temperature phenomena in astrophysics.

Main constraints for the size of ED can be obtained from observations of neutrinos from SN1987A [574–576], γ emission of neutron stars [577], overcooling of early universe [578], and relic photons [579].

SN1987A Some constraints that observations of neutrinos from SN1987A can give were discussed in Section 2.10.2. Similar arguments can be applied to any neutrino-like object, that is, excessive production of weakly interacting particles would overcool thermal states of the hot core. The fact that the neutrinos were observed over an interval of several seconds was used to limit the total amount of cooling and thus the strength of the axion coupling (see Section 7.4.1). The KK graviton would have similar effects. The graviton coupling to nonrelativistic nucleons has an extra reduction factor $\sim T^2/\overline{M}_D^2$, and the thermally averaged cross section for the graviton Bremsstrahlung is estimated roughly as [547, 574]

$$<\sigma v>\sim (30\ \text{mb})\left(\frac{T}{\overline{M}_D}\right)^{\delta+2} \tag{6.156}$$

During the supernova collapse, as much as $\sim 10^{53}$ erg of energy is released in a few seconds, which agrees with the SM of the stellar evolution. Therefore, we require that the graviton luminosity is less than 10^{53} erg/s $\sim (10^{16}\text{GeV})^2$. The graviton luminosity from the nucleon–nucleon Bremsstrahlung is given by [574]

$$L_G \sim (30\ \text{mb})M_{\text{core}}\frac{n_N^2}{\rho}T\left(\frac{T}{\overline{M}_D}\right)^{\delta+2} \tag{6.157}$$

where $M_{\text{core}} \sim 1.6M_\odot \sim 10^{57}$ GeV is the core mass of the supernovae, $n_N \sim 10^{-3}$ GeV^3 is the nucleon number density in the core, $\rho \sim 10^{-3}$ GeV^4 is the mass density. One finds

$$\overline{M}_D \sim 10^{\frac{(15-4.5\delta)}{(\delta+2)}}\ \text{TeV} \tag{6.158}$$

A detailed calculation was done in [575], which was shown to lead to

$$\overline{M}_D \gtrsim \begin{cases} 50\ \text{TeV} & \rightarrow \quad R \lesssim 3\times 10^{-4}\ \text{mm} \qquad \delta = 2 \\ 4\ \text{TeV} & \rightarrow \quad R \lesssim 4\times 10^{-7}\ \text{mm} \qquad \delta = 3 \\ 1\ \text{TeV} & \rightarrow \quad R \lesssim 2\times 10^{-8}\ \text{mm} \qquad \delta = 4 \end{cases} \tag{6.159}$$

There is no bound for $\delta > 4$. Notice that these bounds are far more stringent than the present bounds (~ 0.1 mm) obtained by the torsion experiments (see Figure 6.1).

Neutron Star: After a supernova explosion, most of the KK gravitons stay gravitationally trapped in the remnant neutron star. The requirement that neutron stars are not excessively heated by KK particles decaying into photons leads to

$$M_D > 1700\,(76)\ \text{TeV} \quad \text{for } \delta = 2\,(3) \qquad [577, 580] \tag{6.160}$$

Early Universe In the early universe, the temperature is high and the KK gravitons are produced copiously. Part of them escape to the ED, and will cause excess cooling of the hot plasma. The decay rate for a graviton into two photons ($G \rightarrow 2\gamma$) is given by [578]

$$\Gamma_{G\rightarrow\gamma\gamma} = \frac{m_G^3}{80\pi\overline{M}_{\text{Pl}}^2} \tag{6.161}$$

where m_G is the mass of the graviton. It corresponds to a decay time of

$$\tau \sim (10^8\ \text{Gyr})\left(\frac{\text{MeV}}{m_G}\right)^3 \tag{6.162}$$

Equation (6.162) means that, once a graviton is produced, it decouples from the thermal plasma and does not decay for a long time.

If the cooling rate due to the graviton exceeds that of Hubble cooling, the Hubble expansion will no longer be ordinary as we understand it (see Chapter 8). We can define a normalcy temperature T_* below which the Hubble expansion proceeds normally. It is also the maximum reheating temperature after the cosmic inflation (see Figure 8.8) because, above it, exotic consequences of the ED are effective and evolution history of the universe will become nonstandard. The normalcy temperature can be determined by equating the Hubble cooling rate and that of the graviton cooling.

The standard Hubble cooling rate can be determined as follows: using the Friedmann equation $H^2 \sim G_N \rho$ (ρ is the cosmic energy density. See Eq. (8.10)) and $\rho \sim T^4$ (see Eq. (8.39), cooling by Hubble expansion is roughly

$$\frac{d\rho}{dt} \overset{(8.19)}{\sim} -3H\rho \overset{(8.10)}{\sim} -\frac{T^6}{\overline{M}_{\text{Pl}}} \tag{6.163}$$

The graviton production cross section is proportional to $1/\overline{M}_D^{\,\delta+2}$ [see Eq. (6.155)]. The graviton's evaporation cooling rate can be determined by dimensional analysis to be

$$\left.\frac{d\rho}{dt}\right|_{\text{grav}} \sim -\frac{T^{\delta+7}}{\overline{M}_D^{\,\delta+2}} \tag{6.164}$$

Comparing Eq. (6.163) and (6.164), we obtain a constraint for the normalcy temperature [574]

$$T_* < 10^{\frac{6\delta-9}{\delta+2}}\ \text{MeV} \times \left(\frac{\overline{M}_D}{1\ \text{TeV}}\right)^{\frac{\delta+2}{\delta+1}} \simeq \begin{cases} 10\,\text{MeV} & \delta = 2 \\ \quad\vdots & \\ 10\,\text{GeV} & \delta = 6 \end{cases} \tag{6.165}$$

The astrophysical constraints prefer $\overline{M}_D \sim 10$ TeV, in which case T_* moves up by a factor 10. We need at least $T_* > 1$ MeV in order for ordinary BBN (big-bang nucleosynthesis; see Section 8.6.5) to be recovered. On the other hand, if T_* is too large, then copious production of the bulk graviton by the SM particles can alter the standard cosmology in unacceptable ways. $T_* = 100$ MeV may be considered safe.

Cosmic Diffuse Gamma Rays: If the graviton's mass is as large as 100 MeV, Eq. (6.162) tells that they are decaying at present time contributing to the cosmic diffuse γ (CDG) ray background. Distortion of the CDG spectrum (see Figure 6.14) due to the $G_{\text{KK}} \to \gamma\gamma$ decays was considered by [579] and gives $\overline{M}_D > 100$ TeV for $\delta = 2$ and $\overline{M}_D > 5$ TeV for $\delta = 3$.

6.8.2
Collider Experiments on ADD Models

Graviton Production in Large Extra Dimension Scenario We consider physical processes for probing EDs relevant to collider experiments. Here we first treat

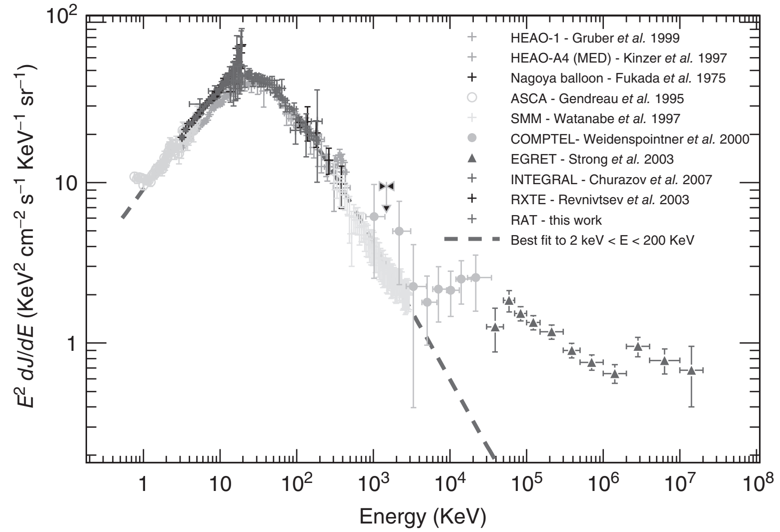


Figure 6.14 Multiwavelength spectrum from X-rays to γ-rays. The dashed line is the best fit to 2 keV $< E <$ 2000 keV. (Reproduced with permission of [581].)

graviton productions in the ADD scenario, that is, we assume that only the graviton propagates in the bulk and the SM particles are confined in the brane. As Eq. (6.154) shows, mass splittings are less than experimental resolutions and from practical point of view can be regarded as continuous. Coupling strength of the graviton to the SM particles is weak ($1/\overline{M}_{\rm Pl}^2$) but is compensated by the large phase space volume $\sim \overline{M}_{\rm Pl}^2$. The number of KK modes between $|n|$ and $|n| + dn$ is given by [467]

$$dN = S_{\delta-1}|n|^{\delta-1}dn, \qquad S_{\delta-1} = \frac{2\pi^{\frac{\delta}{2}}}{\Gamma(\frac{\delta}{2})} \tag{6.166}$$

where $S_{\delta-1}$ is the total solid angle in δ dimension. Using $m = |n|/R$ and Eq. (6.154) for the expression of R, we have

$$dN = S_{\delta-1}\frac{\overline{M}_{\rm Pl}^2}{M_D^{2+\delta}}m^{\delta-1}dm \quad {}^{22)} \tag{6.167}$$

---

22) In the literature, cross sections are often given in terms of M_D [467]. M_D is related to $\overline{M}_D$ by the relation

$$\begin{aligned} \frac{1}{8\pi G_N} &= \overline{M}_{\rm Pl}^2 = \overline{M}_D^{2+\delta}(2\pi R)^\delta \\ &= M_D^{2+\delta}R^\delta \quad \to \quad \overline{M}_D = (2\pi)^{-\frac{\delta}{\delta+2}}M_D \sim (0.4 - 0.25)\,M_D \end{aligned} \tag{6.168}$$

Although there is not much difference between the two for handwaving arguments, we follow the original author's convention.

Hence, the differential cross section for inclusive graviton production can be expressed as

$$\frac{d^2\sigma}{dtdm} = S_{\delta-1}\frac{\overline{M}^2_{\rm Pl}}{M_D^{2+\delta}}m^{\delta-1}\frac{d\sigma_m}{dt} \tag{6.169}$$

where $d\sigma_m/dt$ is the differential cross section for producing a single KK graviton of mass m. The coupling strength of the graviton with matter is given in Eq. (6.99a). The cross section for producing a graviton and a photon in a fermion-anti-fermion collision is given by

$$\frac{d\sigma_m}{dt}(f\overline{f} \to \gamma G) = \frac{\alpha Q_f^2}{16N_f}\frac{1}{s\overline{M}^2_{\rm Pl}}F_1\left(\tfrac{t}{s}, \tfrac{m^2}{s}\right) \tag{6.170}$$

Here, Q_f and N_f are the electric charge and number of colors of the fermion f, the Mandelstam variable t in Eq. (6.170) is defined as $t = (p_f - p_G)^2$, and F_1 is a polynomial given in Appendix C.

Graviton production at the electron collider: To get a feeling for the size of the signal, consider a process where $f = e$, p at 1 TeV center-of-mass energy. The result is shown in Figure 6.15.

The signal for the graviton production would be an energetic photon and missing energy. Examples of experimental searches at LEPII are given in Figure 6.16.

The ADL (ALEPH, DELPHI, L3) combined limits are given in Table 6.2

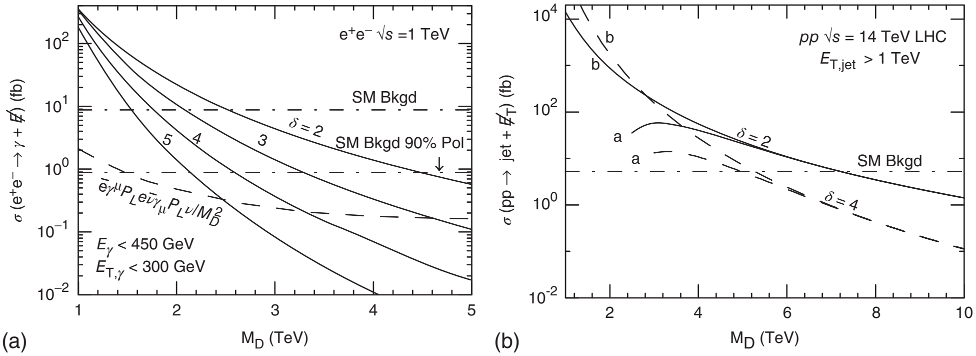


Figure 6.15 (a) Total $e^+e^- \to \gamma + G$ cross section at 1 TeV center-of-mass energy for δ between 2 and 6. The cross section for graviton production is presented as solid lines for various numbers of the extra dimension ($\delta = 2, 3, 4, 5$). The SM background for unpolarized beams is given by the upper dash-dotted line, and the background with 90% polarization is given by the lower dash-dotted line. The signal and background are computed with the requirement $E_\gamma < 450$ GeV in order to eliminate the $\gamma Z \to \gamma\nu\overline{\nu}$ contribution to the background. The dashed line is the Standard Model background-subtracted signal from a representative dimension-6 operator. (b) The total jet + nothing cross section versus M_D at the LHC integrated for all $E_{\rm T\ jet} > 1$ TeV with the requirement that the rapidity $|\eta_{\rm jet}| < 3.0$. The SM background is the dash-dotted line, and the signal is plotted as solid and dashed lines for $\delta = 2$ and 4 extra dimensions. The "a" ("b") lines are constructed by integrating the cross section over $\hat{s} < M_D^2$ (all $\hat{s}$). (Reproduced with permission of [467])

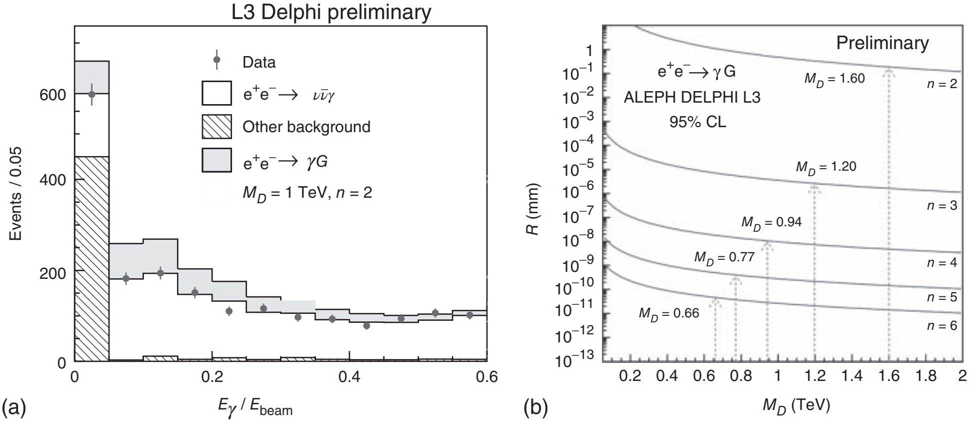


Figure 6.16 (a) Distribution of the ratio of the photon energy to the beam energy for the single-photon events selected by DELPHI and L3, together with the SM prediction. Expected signal from the reaction $e^-e^+ \to \gamma G$ is also shown for $M_D = 1$ TeV and $n = 2$. (b) The radii of the extra dimensions, R, as functions of the gravity scale M_D, for $n = 2\sim6$. Arrows indicate the obtained upper limits on R. (Reproduced with permission of [582].)

Table 6.2 ADL combined limit for MD.

δ	2	3	4	5	6
M_D (TeV)	>1.6	>1.2	> 0.94	>0.77	>0.66

Graviton production in hadron colliders: Feynman diagrams relevant for hadron colliders, that is, for the parton process producing a graviton plus jets are shown in Figure 6.17. The cross sections are given by [467, 583].

$$\frac{d\sigma_m}{dt}(q\bar{q} \to gG) = \frac{\alpha_s}{36}\frac{1}{s\overline{M}_{\rm Pl}^2}F_1\left(\frac{t}{s}, \frac{m^2}{s}\right) \tag{6.171a}$$

$$\frac{d\sigma_m}{dt}(qg \to qG) = \frac{\alpha_s}{96}\frac{1}{s\overline{M}_{\rm Pl}^2}F_2\left(\frac{t}{s}, \frac{m^2}{s}\right) \tag{6.171b}$$

$$\frac{d\sigma_m}{dt}(gg \to gG) = \frac{3\alpha_s}{16}\frac{1}{s\overline{M}_{\rm Pl}^2}F_3\left(\frac{t}{s}, \frac{m^2}{s}\right) \tag{6.171c}$$

where F_1, F_2, F_3 are given in Appendix C. All the cross sections given above are valid below the scale M_D because they were based on the low-energy effective theory. In the hadron collider, one looks for mono jet plus missing energy where the jet may originate from energetic quarks or gluons. To eliminate the SM background such as $Z \to \nu\bar{\nu}$, one has to apply an energy cut in the several hundred GeV to TeV range. An example of such calculations is shown in Figure 6.15b.

Figure 6.18a shows a CMS data on monojet plus missing energy $E_{\rm T}^{\rm miss}$. The expected number of events for the SM and ADD large extra dimension model with

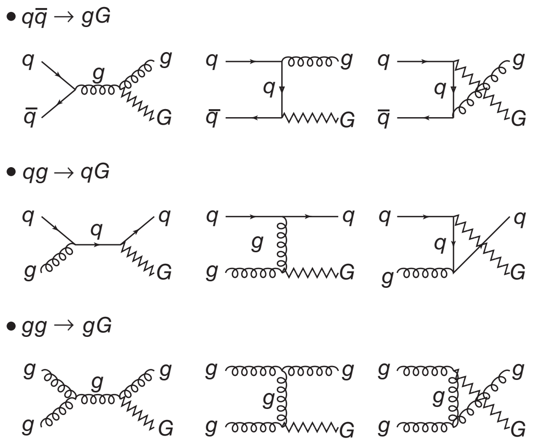


Figure 6.17 Feynman diagrams for $q\bar{q} \to gG$, $qg \to qG$, $gg \to gG$.

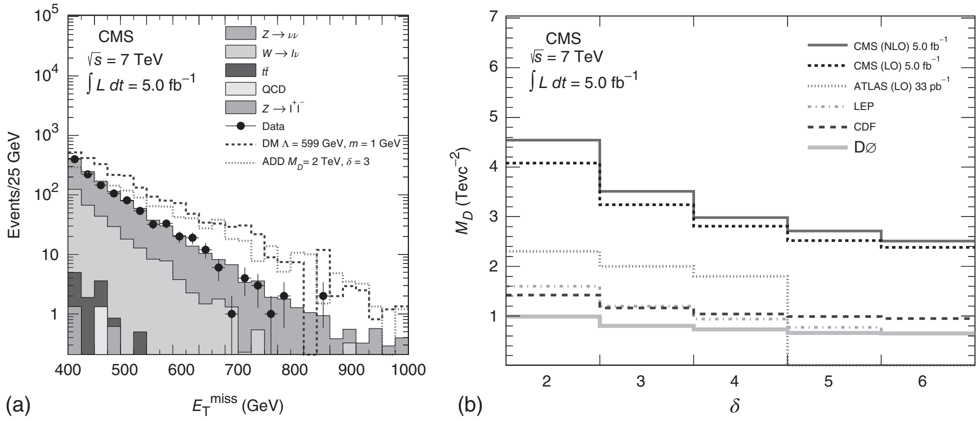


Figure 6.18 Search for the graviton and the dark matter (DM) in "monojet + $\not{E}_T$" samples (a) The distribution of E_T^{miss} for data (black full points with error bars) and simulation (histograms) for $E_T^{\text{miss}} > 350$ GeV c^{-1} after the full event selection criteria are applied. The $Z(\nu\nu)$ + jets and W + jets backgrounds are normalized to their estimates from data. An example of the DM signal (for axial-vector couplings and $M_\chi = 1$ GeV c^{-2}) is shown as a dashed histogram, and an ADD signal (with $M_D = 2$ TeV, $\delta = 3$) is shown as a dotted histogram. (b) Comparison of lower limits on M_D versus the number of extra dimensions with ATLAS [585, 586], LEP [587–590], CDF [591], and D0 [592] (Reproduced with permission of [584].)

$M_D = 2$ TeV, $\delta = 3$ are also drawn as a function of the E_T^{miss}. No excess beyond the SM expectation was observed, and values for M_D smaller than 4.54, 3.51, 2.98, and 2.51 TeV for a number of extra dimensions $\delta = 2, 3, 4, 5$, and 6 were obtained [584]. The ATLAS group obtained a similar result [585]. Figure 6.18b shows a comparison of the lower limits on M_D versus δ for various experiments.

Virtual Graviton Exchange in Large Extra Dimension Scenario We now consider an effect of the virtual graviton exchange at tree level in scattering processes. These processes will be sensitive to the ultraviolet cutoff M_S, which is expected to be $\sim M_D$ and more likely $< M_D$. For simplicity, we consider the case of pure s-channel exchange. The discussion of t and u channel exchange is similar. The scattering amplitude in momentum space of the graviton-mediated process is given by [467]

$$\mathcal{A} = \frac{1}{\overline{M}_{\rm Pl}^2} \sum_n \left[T^{\mu\nu} \frac{P_{\mu\nu\alpha\beta}}{s - m_n^2} T^{\alpha\beta} + \left(\frac{\kappa}{3}\right)^2 \frac{T^\mu_\mu T^\nu_\nu}{s - m_n^2} \right] \equiv \mathcal{S}\mathcal{T} \tag{6.172}$$

$$\begin{aligned} P_{\mu\nu\alpha\beta} &= \frac{1}{2}(\eta_{\mu\alpha}\eta_{\nu\beta} + \eta_{\mu\beta}\eta_{\nu\alpha} - \eta_{\mu\nu}\eta_{\alpha\beta}) \\ &\quad - \frac{1}{2m_n^2}(\eta_{\mu\alpha}k_\nu k_\beta + \eta_{\nu\beta}k_\mu k_\alpha + \eta_{\mu\beta}k_\nu k_\alpha + \eta_{\nu\alpha}k_\mu k_\beta) \\ &\quad + \frac{1}{6}\left(\eta_{\mu\nu} + \frac{2}{m_n^2}k_\mu k_\nu\right)\left(\eta_{\alpha\beta} + \frac{2}{m_n^2}k_\alpha k_\beta\right) \end{aligned} \tag{6.173}$$

$$\mathcal{S}(s) \equiv \frac{1}{\overline{M}_{\rm Pl}^2} \sum_n \frac{1}{s - m_n^2}, \qquad \mathcal{T} \equiv T_{\mu\nu}T^{\mu\nu} - \frac{1}{\delta + 2} T^\mu_\mu T^\nu_\nu \tag{6.174}$$

Here, $T^{\mu\nu}$ is the energy–momentum tensor, k is the transferred momentum, and κ is defined in Eq. (6.97). $\sum_n$ represents the sum over all KK modes. As the operator $\mathcal{T}$ is universal to all the KK modes, the summation can be carried out. It yields a diverging integral that can be regularized by introducing an effective cut-off factor M_S. One expects $M_S \sim M_D$, and it can be considered as the effective Planck scale. The presence of ultraviolet divergence in tree-level processes is related to existence of infinite tower of KK modes. For $\delta = 2$ case, there is no power divergence and

$$\mathcal{A} = \begin{cases} -\dfrac{1}{M_D^{\delta+2}} \dfrac{S_{\delta-1}}{2} \left(\ln \dfrac{s}{\mu^2} + i\pi \right) \mathcal{T} & \delta = 2 \\ \dfrac{4\pi}{M_S^4} \mathcal{T} & \delta > 2 \end{cases} \tag{6.175}$$

Here, $S_{\delta-1}$ is the solid angle defined in Eq. (6.166) and μ is the subtraction mass. Using Eq. (6.175), one can calculate cross sections for various processes relevant to collider experiments. For instance, the cross section with two photons in the final state is

$$\frac{d\sigma}{dt}(f\overline{f} \to \gamma\gamma) = \frac{\pi}{N_f s^2} \left[\alpha Q_f^2 G_1\left(\frac{t}{s}\right) + \frac{s^2}{M_S^4} G_2\left(\frac{t}{s}\right) \right]^2 \tag{6.176a}$$

$$\frac{d\sigma}{dt}(gg \to \gamma\gamma) = \frac{\pi}{512} \frac{s^2}{M_S^8} G_3\left(\frac{t}{s}\right) \tag{6.176b}$$

where $G_1 \sim G_3$ are given in Appendix C [467]. A similar expression can be obtained for $f\overline{f} \to f\overline{f}$. A generic form often used for data fitting for the sum of the SM and

KK processes can be written as [593]

$$\frac{d^2\sigma}{d\cos\theta^* dM} = f_{\mathrm{SM}} + f_{\mathrm{int}}\left(\frac{F}{M_S^4}\right) + f_{\mathrm{KK}}\left(\frac{F}{M_S^4}\right)^2$$
$$F = \begin{cases} 1 & \text{GRW [467]} \\ \begin{cases} \ln\frac{M_S^2}{M^2} & \delta = 2 \\ \frac{2}{\delta-2} & \delta > 2 \end{cases} & \text{HLZ [594]} \\ \dfrac{2\lambda}{\pi} = \pm\dfrac{2}{\pi} & \text{Hewett [595]} \end{cases} \tag{6.177}$$

6.8.3
TeV^{-1} Extra Dimension Model:

ADD model: An example of data obtained by the ATLAS group for the production of 2γ as a function of the invariant mass $M_{\gamma\gamma}$ is shown in Figure 6.19 [596].

The data were interpreted for the ADD and the RS models using the formula in Eq. (6.177). A similar result is obtained by the CMS group [597], and their resultant 95% CL lower limits on M_S are shown in Table 6.3.

The K factor reflects theoretical uncertainties in the normalization of the cross section.

RS model: For the TeV^{-1} ED signals, we can also look for resonances produced by the graviton or other KK excited states. As described in Section 6.6.5, KK gravitons could show up as resonances in $\gamma\gamma$, $l\bar{l}$ or two jets. An example of data for production

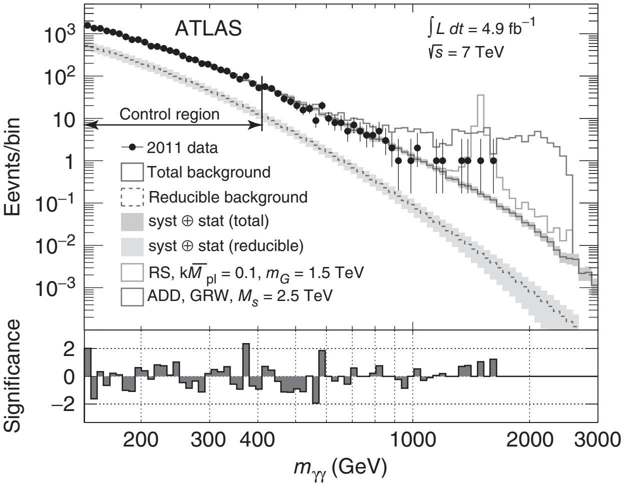


Figure 6.19 Observed invariant mass distribution of diphoton events. Superimposed are the SM background expectation and the expected signals for two examples of RS and ADD models (Reproduced with permission of [596, 597].) Please find a color version of this figure on the color plates.)

Table 6.3 The 95% CL lower limits obtained by CMS on M_S (in TeV) in the GRW [467], Hewett [595], and HLZ [594] conventions for two values of the ADD signal K factors 1.0 and 1.6 ± 0.1. All limits are computed with a signal cross section truncated to zero for $\sqrt{\hat{s}} > M_S$, where $\sqrt{\hat{s}}$ is the center-of-mass energy of the partonic collision. The limits are presented for both positive and negative interference in the Hewett convention and for δ (denoted as n_{ED}) = 2–7 in the HLZ convention. The median expected lower limits are given in parentheses. Reproduced with permission of [597].

	GRW	Hewett		HLZ					
K		Positive	Negative	$n_{ED} = 2$	$n_{ED} = 3$	$n_{ED} = 4$	$n_{ED} = 5$	$n_{ED} = 6$	$n_{ED} = 7$
1.0	2.94	2.63	2.28	3.29	3.50	2.94	2.66	2.47	2.34
	(2.99)	(2.67)	(2.31)	(3.37)	(3.56)	(2.99)	(2.71)	(2.52)	(2.38)
1.6 ± 0.1	3.18	2.84	2.41	3.68	3.79	3.18	2.88	2.68	2.53
	(3.24)	(2.90)	(2.44)	(3.77)	(3.85)	(3.24)	(2.93)	(2.73)	(2.58)

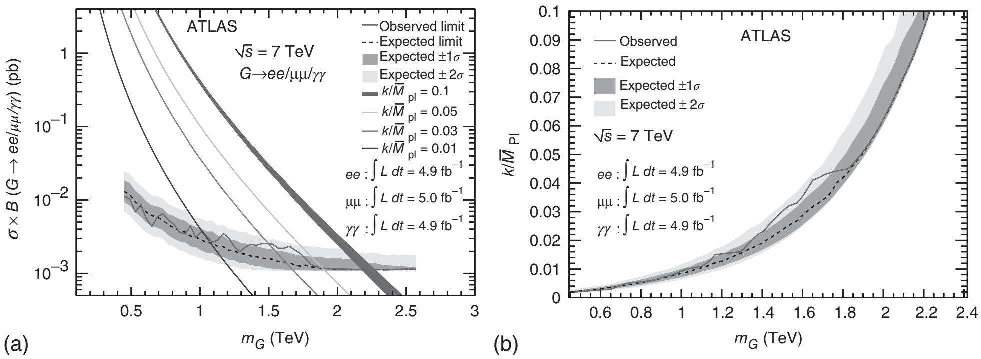


Figure 6.20 (a) ATLAS expected and observed 95% CL limits from combination of $G \to \gamma\gamma/ee/\mu\mu$ channels on $\sigma \times B$, the product of the RS graviton production cross section and the branching ratio for graviton decay via $G \to \gamma\gamma/ee/\mu\mu$, as a function of the graviton mass. The thickness of the theory curve for $k/\overline{M}_{\text{Pl}} = 0.1$ illustrates the theoretical uncertainties due to the PDFs (parton distribution functions) expressed at 90% CL. For $k/\overline{M}_{\text{Pl}} = 0.1$, the lower limit on m_G is ~ 2.23 TeV. (b) The RS interpretation of the data on the plane of the graviton mass versus $k/\overline{M}_{\text{Pl}}$. The region above the curve is excluded at 95% CL. (Reproduced with permission of [596].)

of 2γ as a function of the invariant mass $M_{\gamma\gamma}$ is shown in Figure 6.19. A hypothetical RS graviton signal is overlayed on the data.

They also looked at two-lepton spectra and found that the data agreed quite well with the SM expectations. They obtained upper limits on the cross section × branching ratio. The result is plotted in Figure 6.20a together with RS model expectation for various values of the parameter $k/\overline{M}_{\text{Pl}}$.

The excluded parameter region on the m_G (graviton mass) versus $k/\overline{M}_{\rm Pl}$ is plotted in Figure 6.20b which is to be compared with Figure 6.12.

UED Model: CMS looked for a heavy W boson (W'), which decays via the same decay channel as the SM W ($W' \to l + E_{\rm T}^{\rm miss}$), which has the same coupling as the SM W (referred to as *SSM W = sequential standard model W*). Their event rate of $W' \to \mu\nu$ as a function of the transverse mass M_T is shown in Figure 6.20a. The signal will show up as a bump rather than a peak as a function of the transverse mass M_T. The data was also analyzed using a split UED (sUED) model [598–600]. It is a modified version of the UED model where fermions are confined in a fat brane (see Section 6.3.4). Placing the fermions on the brane modifies the coupling strength of the fermion, allowing more flexibility in fitting the data. As n=2 KK level of $W_{KK}^{n=2}$ can couple directly to the SM particles, it can be analyzed exactly in the same way as the SM-like W'.

No signals for the ED were observed. The upper limit of the W' production cross section is plotted in Figure 6.21(a) for $M_{W'}$=1.3 and 2.3 TeV. Figure 6.21b shows the 95% CLs on the split-UED bulk fermion mass parameter M (denoted as

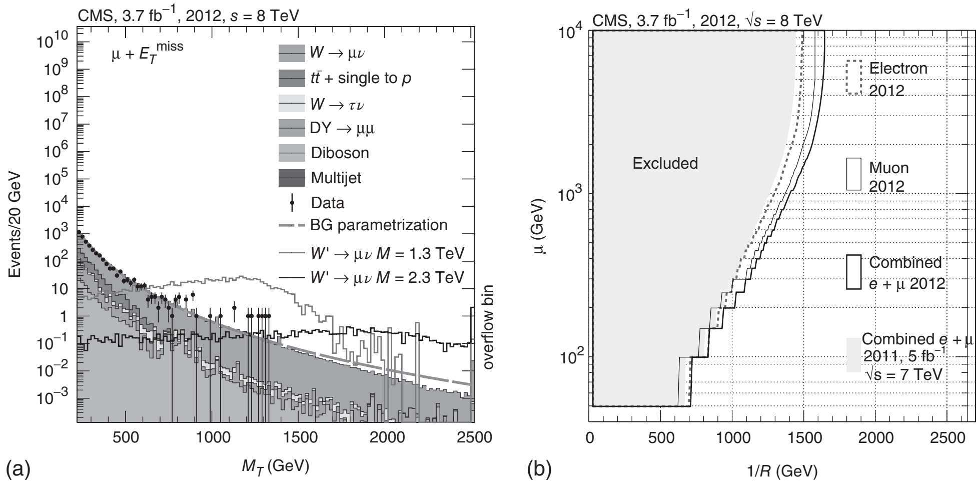


Figure 6.21 (a) Observed lepton-$E_T^{\rm miss}$ transverse mass distributions in the muon channel. The dashed lines show the parametrization of the background. Simulated signal distributions for an SSM W' (which includes $W_{\rm KK}^{n=2}$) are also shown. The simulated background labeled as "diboson" includes WW, ZZ, and WZ contributions. (b) The 95% CLs on the split-UED fermion bulk mass parameters M (denoted as μ in the figure) and $1/R$ derived from the W' mass limits taking into account the corresponding width of the $W_{\rm KK}^{n=2}$. For the 3.7 fb^{-1} of 2012 data, the individual limits for the electron and muon channels are shown together with their combination, improving the excluded parameter space based on the 2011 data shown in grey. (Reproduced with permission of [601].) (Please find a color version of this figure on the color plates.)

μ in the figure)[23] and $1/R$ derived from the W' mass limits taking into account the corresponding width of the $W_{\text{KK}}^{n=2}$. Note that W' appears in many models for the beyond the SM like those discussed in the GUTs in Chapter 3. Therefore, Figure 3.6 and 3.7 can also be used to constrain the ED models.

6.9
Black hole production

A black hole (BH) can be formed when a large mass M is confined within a radius smaller than the Schwarzschild radius $R_{\text{S}} = 2MG_N/c^2$. Naively, BH would only grow by attracting more mass. If quantum effect is taken into account, it was shown that the BH behaves like a thermal black body with the Hawking temperature [602]

$$k_{\text{B}} T_{\text{H}} = \frac{(\delta + 1)}{4\pi R_S} \hbar c \tag{6.178}$$

where k_{B} is the Boltzmann constant. In natural units, one has the fundamental relationship $R_{\text{S}} T_{\text{H}} = (4\pi)^{-1}$. The BH can evaporate by emitting photons, or, if the temperature T_{H} is high enough, massive particles can also be produced in evaporation. Conversely, if two energetic particles collide with the impact parameter (transverse distance) smaller than R_{S}, the BH can be formed with subsequent decays to particles [603–606]. One needs a theory of quantum gravity to calculate the exact cross section, but it is reasonable to assume that the cross section is given by

$$\sigma(E) \sim \pi R_S^2 \tag{6.179}$$

In ED, there is a possibility that the fundamental coupling constant of the gravity, and hence the Schwarzschild radius, could be large. It has opened a possibility to produce microscopic BHs by an accelerator. We use the Schwarzschild radius in the ED

$$R_{\text{S}} = \frac{k(\delta)}{M_{\text{D}}} \left[\frac{M_{\text{BH}}}{M_{\text{D}}} \right]^{\frac{1}{(\delta+1)}}, \qquad k(\delta) = \left(2^{\delta} \pi^{\frac{\delta-3}{2}} \frac{\Gamma\left(\frac{\delta+3}{2}\right)}{\delta + 2} \right)^{\frac{1}{(\delta+1)}} \tag{6.180}$$

as given by [8, 607, 608]. The present constraint Eq. (6.2b) on the size of the LED, $R = (M_{\text{Pl}}^2/M_{\text{D}}^{\delta+2})^{1/\delta} < 0.1$ mm, translates to

$$\begin{aligned} R_{\text{S}} &< k(\delta) \left(\frac{M_{\text{BH}}}{0.75\ \text{TeV}} \right)^{\frac{1}{(\delta+1)}} 10^{-\frac{2\delta+49}{\delta+1}}\ \text{cm} \\ &\simeq \left(\frac{M_{\text{BH}}}{1\ \text{TeV}} \right)^{\frac{1}{(\delta+1)}} \times \begin{cases} 10^{-26}\ \text{cm} & \delta = 1 \\ 10^{-18}\ \text{cm} & \delta = 2 \\ 10^{-14}\ \text{cm} & \delta = 3 \\ \vdots & \end{cases} \end{aligned} \tag{6.181}$$

---

[23] The KK fermion mass is given by $m_f = \sqrt{M^2 + (n/R)^2}$ [see Eq. (6.48)].

Therefore, the LHC may be able to confine a lump of mass within the Schwarzschild radius if $\delta \geq 2$ for the ADD-type models. For the RS model, there is no such restriction because the size of the ED can be taken to infinity. The cross section in ED for $M_D = 1$ TeV becomes

$$\hat{\sigma}(\hat{s} = M_{\rm BH}^2) = \pi R_S^2 \sim 1{\rm TeV}^{-2} \sim 100\,{\rm pb} \tag{6.182}$$

where $\hat{\sigma}$ is the BH production cross section by partons and $\hat{s}$ is the total energy of the partons that participate in the formation of the BH. Notice that the cross section is comparable with that of the top-quark pair production. The BH production cross section at the hadron collider is given by convoluting the cross section with the parton luminosity.

$$\begin{aligned} \frac{d\sigma({\rm pp} \to {\rm BH} + X)}{dM_{\rm BH}} &= \frac{dL}{dM_{\rm BH}}\,\hat{\sigma}(ab \to {\rm BH})|_{\hat{s}=M_{\rm BH}^2} \\ \frac{dL}{dM_{\rm BH}} &= \frac{2M_{\rm BH}}{s}\sum_{a,b}\int_{\frac{M_{\rm BH}^2}{s}}^{1} \frac{dx_a}{x_a} f_a(x_a) f_b\left(\frac{M_{\rm BH}^2}{s x_a}\right) \end{aligned} \tag{6.183}$$

where $f_{a,b}$ are the parton distribution functions in the nucleon. Once the BH of TeV scale is formed, its temperature also being ~ 1 TeV, it decays with lifetime $10^{-27} - 10^{-25}$ s [603, 604, 608]. Let us assume that the decay can be approximated by that of a classical BH. Its temperature is given by

$$T_H = \frac{1+\delta}{4\pi R_S} \simeq \frac{1+\delta}{M_{\rm BH}^{\frac{1}{(1+\delta)}}} \tag{6.184}$$

and multiplicity by

$$< N_{\rm BH} >= \left\langle \frac{M_{\rm BH}}{2E_{\rm particle}} \right\rangle \sim \left\langle \frac{M_{\rm BH}}{2T_H} \right\rangle \tag{6.185}$$

Notice that the BH radiates mainly into our 3D world. This is because the BH is a point radiator producing *s*-waves, decays isotropically, and couples to all the particles with equal strength, that is, the BH decays to all the particles democratically via Hawking radiation. As there are many more particles in the brane than in the bulk, decays into gravitons are largely suppressed. The BH evaporates into (q and g : leptons : Z and W : ν and Graviton : H : γ) = (72% : 11% : 8% : 6% : 2% : 1%), (hadron : lepton) is (5 : 1) accounting for t, W, Z, and H decays [604]. Microscopic BH properties are reviewed in more detail in [605, 609].

Figure 6.22 shows an effort to produce black holes taking best advantage of the assumed democratic radiation, that is, a measurement of multiplicity distribution of the produced final state as a function of scalar sum of transverse energy S_T.

BH signal events are simulated using the parton-level BlackMax [613] generator, followed by a parton showering fragmentation with PYTHIA [614]. No excess was observed beyond the SM expectation, and model-specific indicative limits are set excluding BHs with masses below ~ 5.5 TeV. The results is insensitive to the number of extra dimension.

Notice that notion of the classical black body radiation is valid for $M_{\rm BH} \gg M_D$. Use of $M_{\rm BH}^{\rm min}$ rather than $M_{\rm BH}$ in the analysis reflects this fact.

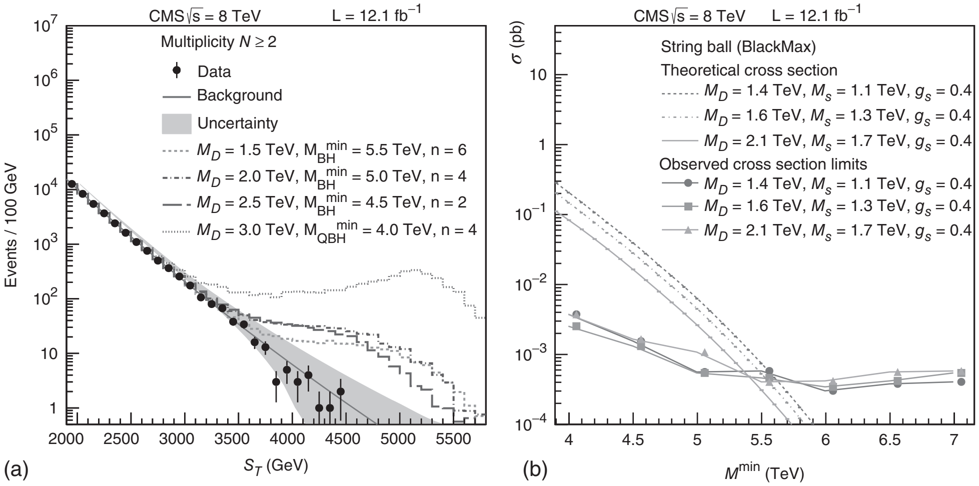


Figure 6.22 (a) Distribution of the total transverse energy S_T with multiplicity $N \geq 2$ objects (photons, electrons, muons, or jets) in the final state. Observed data are depicted as points with error bars; solid line with a shaded band is the background prediction and its systematic uncertainty. Also shown are the expected semiclassical black hole signals for four parameter sets of the BlackMax nonrotating black hole model. Here, $M_{\mathrm{BH}}^{\mathrm{min}}$ is the minimum black hole mass, M_{D} is the multidimensional Planck scale, and n (δ in the text) is the number of extra dimensions. (b) The cross section upper limits at 95% CL from the counting experiments optimized for various black hole parameter sets (solid lines) compared with signal production cross sections from the BlackMax generator (dashed lines) as a function of minimum black hole mass (Reproduced with permission of [610–612].) (Please find a color version of this figure on the color plates.)

Quantum black hole (QBH): The assumption of the Hawking radiation may not be justified as it is only valid for $T \gg M_{\mathrm{BH}}$. As the BH mass approaches the Planck scale M_{D}, quantum effects become important. Despite lack of a complete theory of quantum gravity, it is still possible to gain insight into the nature of QBHs based on some fundamental principles and a few assumptions [615–617]. Two central assumptions are that formation of QBHs conserves QCD and $U(1)$ charges as local gauge symmetries are not violated by gravity, and that it takes place in a limited region of spacetime. The latter assumption provides a cutoff for calculations. Then, the formation of the BH might be viewed as a strongly coupled resonance or a gravitationally bound state. Such objects evaporate faster than they thermalize, resulting in nonthermal decays into a small number of SM particles. The decay width will be narrower and rather insensitive to the number of ED.

One estimate [616] gives $\Gamma \simeq (1/4\pi)M_{\mathrm{BH}}^5/M_D^4$, which is of the order of 80 GeV for a mass of 1 TeV. The dijet production cross section is a powerful tool to detect a new physics. The CMS group analyzed the data for possible detection of new particles including excited quark (q^*), string resonances (S), as well as the QBH. Figure 6.23(a) shows CMS data on the dijet production cross section as a function of dijet mass.

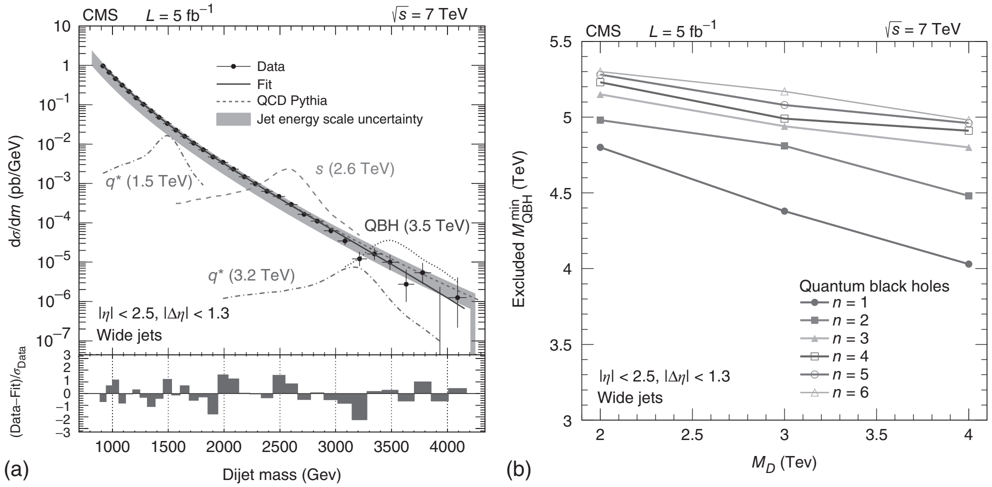


Figure 6.23 (a) Inclusive dijet mass spectrum and predictions for QCD (short-dashed), excited quarks (q^*), string resonances (S), and quantum black holes (QBHs). The shaded band shows the contribution from the systematic uncertainty in the jet energy scale. (b) Observed 95% CL lower limits on the minimum mass of QBHs as a function of the quantum gravity scale M_D for the number of extra dimensions δ (n in the figure) of 1 (RS model) and 2–6 (ADD model) (Reproduced with permission of [384].)

Here we concentrate on the QBHs. As the data was consistent with the SM predictions, model-dependent upper limits on the minimum value of the QBHs as a function of the Planck scale in ED were obtained. The results are similar to those obtained by the multiplicity distributions. Lower limits of $M_{\text{BH}}^{\text{min}}$ of ∼ 4–5 TeV were obtained for both RS- and ADD-type EDs (Figure 6.23(b)).

7
Axion

“Axion” is a name given to a particle that was predicted in 1977 to solve the so-called strong CP problem [618]. It is very light (tens of eV to 10^{-6} eV), interacting very weakly with matter and has not been discovered to this day despite the countless number of efforts made in the search for its existence. Accumulation of data has narrowed the allowed range of the axion parameters (mass and coupling strength) to a limited region. Behind the long efforts for its searches lie many outstanding theoretical issues such as the chiral anomaly, the θ vacuum, and the instanton. Besides, the axion itself is a promising candidate for the dark matter and occupies an important status in cosmology.

The θ vacuum and instanton are results of nonperturbative treatments of the QCD and reveal a completely different aspect of the field theory. Consequently, we take a roundabout route to attack the axion problem. We first discuss the physical idea of the soliton; introduce a topologically conserved number (the winding number), the instanton as a soliton in four-dimensional Euclidean space, the relationship between the θ vacuum and the chiral anomaly; and see how they are connected to the instanton. All these themes are interesting theoretically, but have no direct relationship with the axion phenomenology. Therefore, readers who are only interested in the axion phenomenology may skip Sections 7.1 and 7.2 and start from Section 7.3.

7.1
Soliton

7.1.1
Kink

In simple words, the soliton (solitary wave) is a wave localized in space and stable in time, and appears as a nonperturbative solution to a nonlinear wave equation [619–622]. A conventional wave that obeys a linear wave equation has dispersion, which means the relation $\omega^2 = k^2 + m^2$, where ω, k, m denote the frequency, wave number, and a constant. Its wave form dissociates as time passes, but the soliton is intrinsically stable. In order to understand a physical picture of the soliton, let us start with a classical scalar field equation in the 1 + 1 dimension.

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

Conditions for the Existence of the Soliton We consider the following Lagrangian density:

$$\mathscr{L} = \frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi) = \frac{1}{2}\left(\frac{\partial\phi}{\partial t}\right)^2 - \frac{1}{2}\left(\frac{\partial\phi}{\partial x}\right)^2 - V(\phi) \tag{7.1}$$

where we use $x^\mu = (t, x)$, and $V(\phi)$ represents a nonlinear potential. It has at least one minimum, and a constant in it is adjusted to assure that the minimum satisfies $V_{\text{min}} = 0$. The equation of motion does not necessarily have a static solution and there is a possibility of having a nonstatic solution. Here, for simplicity, we limit our discussion to a static solution. When a static solution $\phi = f(x)$ is found, the time-dependent solution can be obtained by Lorentz transformation. That is

$$\phi = f(\xi), \quad \xi = \gamma(x - vt) \tag{7.2}$$

is the desired solution. $\gamma = (1 - v^2)^{1/2}$ is the Lorentz γ factor. According to the Hamilton principle, the equation of motion for the static solution is written as

$$\frac{\partial^2\phi}{\partial x^2} = \frac{\partial V}{\partial\phi} \tag{7.3}$$

The energy density of the field is expressed as

$$\mathscr{H}(\phi) = \frac{1}{2}\left(\frac{\partial\phi}{\partial x}\right)^2 + V(\phi) \tag{7.4}$$

The total energy is given by

$$E(\phi) = \int_{-\infty}^{+\infty} dx\, \mathscr{H}[\phi(x)] = 2\int_{-\infty}^{+\infty} dx\, V[\phi(x)] \tag{7.5}$$

Justification of the second equality will be given later in (7.7).

We will consider a condition for the potential $V(\phi)$ to have soliton solutions. When a soliton solution exists, it should have a finite energy and hence its energy density has to be localized in space. Looking at Eqs. (7.4) and (7.5), the field ϕ has to satisfy a relation $\partial\phi/\partial x = V(\phi) = 0$ at $x \to \pm\infty$. We notice that the equation of motion (7.3) reduces to that of Newtonian mechanics with $m = 1$ if we change $x \to t$, $\phi \to x$, and $V \to -V$. When the potential V has a single minimum like one in Figure 7.1a, the inverted potential looks like the one in Figure 7.1b. The boundary condition requires that $\phi = \phi_0$ at $x = \pm\infty$. Suppose one starts from $x = -\infty$. If ϕ shifts from the position $\phi = \phi_0$, the field ϕ rolls down the potential and cannot avoid reaching $\phi = +\infty$ or $-\infty$. Consequently, there is no other solution but to stay steady at $\phi = \phi_0$.

What will happen if there are more than two minima in the potential? For instance, let us consider a W-shaped potential (see Figure 7.2a).

$$V(\phi) = \frac{\lambda}{4}(|\phi|^2 - v^2)^2 \tag{7.6}$$

The inverted potential looks like Figure 7.2b. Assuming $\lim_{x\to-\infty}\phi(x) = -v$, there is a solution for ϕ to move to the right and reach $\lim_{x\to+\infty}\phi(x) = +v$. The inverse solution $\phi = v \to -v$ exists, too. In this case, the field ϕ reaches v

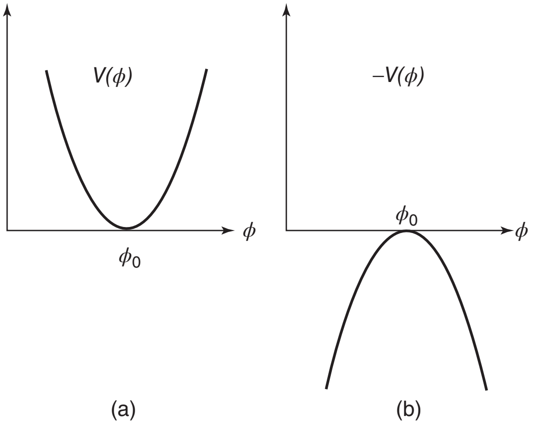


Figure 7.1 Normal (a) and inverted (b) potentials with one minimum.

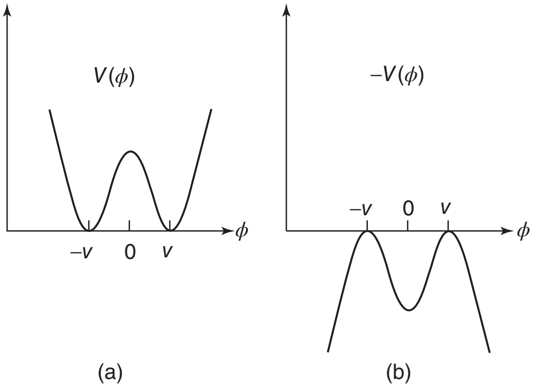


Figure 7.2 Potential shape of a spontaneously broken symmetry (a) and its inverted potential. (b) The W-shaped potential in its rotated form is often referred to as the *wine bottle* or *Mexican hat* potential.

asymptotically and never goes back to $\phi = -v$ nor goes beyond $\phi = v$. The above consideration leads us to conclude that, for the soliton solution to exist, at least two minima are required. It also suggests that the existence of the soliton is closely connected with the spontaneous symmetry breakdown.

Given such boundary conditions, a general solution to Eq. (7.3) can be obtained immediately. Multiplying $\partial\phi/\partial x$ on both sides and integrating, we obtain

$$\frac{1}{2}\left(\frac{\partial\phi}{\partial x}\right)^2 = V(\phi) \tag{7.7}$$

which gives the promised proof for the second equality of Eq. (7.5). As $\partial\phi/\partial x = V(\phi) = 0$ as $x \to \pm\infty$, the integration constants have to vanish at both ends. Hence

$$x = x_0 \pm \int_{\phi(x_0)}^{\phi(x)} \frac{d\phi}{\sqrt{2V(\phi)}} \tag{7.8}$$

Using Eq. (7.8), we can show that the soliton solutions for the potential Eq. (7.6) are given by

$$\begin{aligned} \phi(x) &= \pm v \tanh\left[\frac{m}{\sqrt{2}}(x - x_0)\right], \quad v^2 = \frac{m^2}{\lambda} \\ E &= \frac{2\sqrt{2}m^3}{3\lambda} = \frac{2\sqrt{2}}{3}\sqrt{\lambda}v^3 \end{aligned} \tag{7.9}$$

The two solutions, which are referred to as the *kink* and *anti-kink*, and their energy density are given in Figure 7.3. One sees that the energy is localized at $x = x_0$ with its spread given by $1/m$.

Sine–Gordon Equation As an example of the infinite number of degenerate vacua, we consider the Sine–Gordon equation, which is given by

$$\frac{\partial^2\phi}{\partial t^2} - \frac{\partial^2\phi}{\partial x^2} + a \sin b\phi = 0 \tag{7.10}$$

The potential to give the Sine–Gordon equation is expressed as

$$V(\phi) = \frac{a}{b}(1 - \cos b\phi) \tag{7.11}$$

It has the shape depicted in Figure 7.4a and has minima at

$$\phi = \frac{2\pi n}{b}, \quad n = 0, \pm 1, \pm 2, \cdots \tag{7.12}$$

which represents an infinite number of degenerate vacua. The solution to the Sine–Gordon equation can be obtained by substituting Eq. (7.11) into Eq. (7.8).

$$\phi(x) = \pm\frac{4}{b}\tan^{-1}\left[\exp\{\pm\sqrt{ab}(x - x_0)\}\right] \tag{7.13}$$

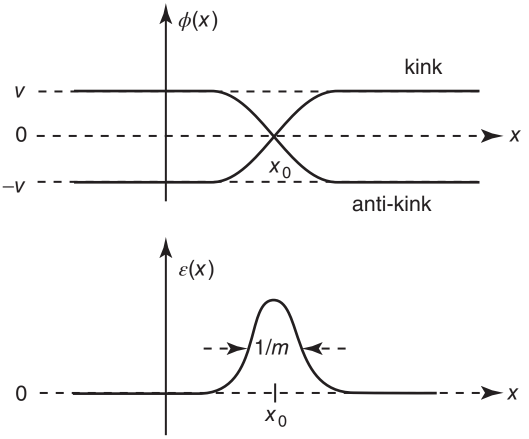


Figure 7.3 Kink/anti-kink solutions and the associated energy density for ϕ^4 theory in (1 + 1) dimensions.

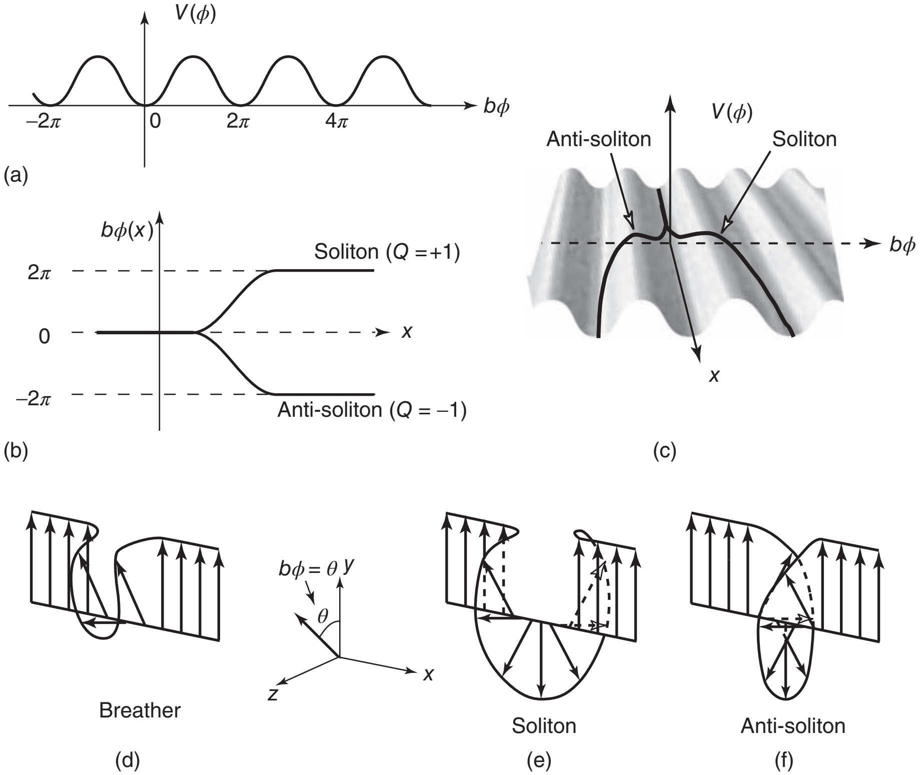

$V(\phi)$
$b\phi$
-2π
0
2π
4π
(a)
$b\phi(x)$
2π
0
-2π
Soliton (Q =+1)
x
Anti-soliton (Q = −1)
(b)
$V(\phi)$
Anti-soliton
Soliton
$b\phi$
x
(c)
$b\phi = \theta$
y
θ
x
z
Breather
(d)
Soliton
(e)
Anti-soliton
(f)

Figure 7.4 (a) Potential of the Sine–Gordon Lagrangian. (b) The soliton ($Q = +1$)/anti-soliton ($Q = -1$) solutions as a function of x. (c) 3D illustration of the Sine–Gordon potential and trajectories of a soliton and an anti-soliton. (d)–(f) Elastic ribbon model for the soliton. Arrows denote the orientation of the ribbon. (d) Breather (e) soliton and (f) anti-soliton have all different boundary conditions and cannot decay to one another. [623]

In order to see the meaning of the constant a, b, we expand the potential in the vicinity of $\phi = 0$:

$$V(\phi) = \frac{ab}{2}\phi^2 - \frac{ab^3}{4!}\phi^4 + \cdots \tag{7.14}$$

If we replace $ab \to m^2$, $ab^3 \to -\lambda$, it is a self-interacting scalar field with mass m and self-coupling strength λ. Now let us consider a solution that satisfies a boundary condition such that it goes to one zero point of the potential as $x \to -\infty$ and another (e.g., at $n = \pm 1$) as $x \to +\infty$ (see Figure 7.4b). By setting $\phi(x) = \pm\pi m/\sqrt{\lambda}$ at $x = x_0$, the solution is given by

$$\phi_{\pm 1}(x) = \pm\frac{4m}{\sqrt{\lambda}} \tan^{-1}[e^{m(x-x_0)}] \tag{7.15}$$

Examples of the solutions and their corresponding potential energy are illustrated in Figure 7.4c as a 3D plot in the $x - \phi$ plane. By looking at the energy density of the soliton, which is similar to the lower figure of Figure 7.3, we may interpret that $x = x_0$ is the location of the soliton and $1/m$ gives its spatial size. ϕ_{+1} and ϕ_{-1} have, in many respects, properties of a particle and an anti-particle, as will be described later. The soliton's energy can be easily calculated to give

$$\begin{aligned} E &= 2\int_{-\infty}^{+\infty} dx\, V[\phi(x)] = \left(\frac{2a}{b}\right)^{\frac{1}{2}} \int_0^{\frac{2\pi}{b}} \left[1-\cos(b\phi)\right]^{\frac{1}{2}} d\phi \\ &= 8\left(\frac{a}{b^3}\right)^{\frac{1}{2}} = \frac{8m^3}{\lambda} \end{aligned} \tag{7.16}$$

The functional form of the energy means that it diverges as the coupling strength λ goes to zero. That is, no matter how small the coupling strength is, there exists the soliton as a nonlinear solution to the equation of the motion. It is not a solution that can be obtained perturbatively. The soliton's stability is based on its connection to two different vacua (see Figure 7.4b,c). Its energy is concentrated in a limited region, but it takes infinite energy to change one solution to another by a continuous deformation.

To understand the reason, let us consider a simple model like a twisted ribbon. One edge of the ribbon is fixed along the x-axis but the other edge is at an angle $\theta = b\phi(x)$ on the $y-z$ plane. The $n = \pm 1$ vacua correspond to the ribbon twisted once counterclockwise [Figure 7.4e] or clockwise [Figure 7.4f] and are called the *kink* and the *anti-kink*. By shaking the ribbon, the twist becomes a traveling wave. Obviously, the wave is stable and cannot decay or transform to another in the untwisted state ($n = 0$ vacuum, see Figure 7.4d). The stability of the soliton stems from its topological structure.

The above statement suggests a conserved quantity Q (topological charge) which is characterized by the difference of integers. Let us consider the following current $J^\mu(t,x)\,(\mu = 0,1)$:

$$J^\mu = \frac{b}{2\pi}\epsilon^{\mu\nu}\partial_\nu\phi \qquad \epsilon^{\mu\nu} = -\epsilon^{\nu\mu}, \quad \epsilon^{01} = 1 \tag{7.17a}$$

$$Q = \int_{-\infty}^{+\infty} J^0 dx = \frac{b}{2\pi}\int_{-\infty}^{+\infty} \frac{\partial\phi}{\partial x} dx = \frac{b}{2\pi}\left[\phi(\infty) - \phi(-\infty)\right] = N \tag{7.17b}$$

It is not the Noether current that can be derived from the symmetry of the Lagrangian. The conservation of the current originates from the difference of the boundary condition, that is, the topology of the field. We understand that the two solutions in Eq. (7.15) correspond to a soliton and an anti-soliton. The stability of the soliton can be rephrased as the conservation of the topological number.

7.1.2
Vortex

We now expand our previous discussion to a complex scalar field in 1 + 2 dimensions. The Lagrangian is given by

$$\mathscr{L} = \left|\frac{\partial\phi}{\partial t}\right|^2 - |\nabla\phi|^2 - V(\phi), \qquad \nabla\phi(x,y) = \left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}\right) \tag{7.18}$$

We adopt $V(\phi)$ given by Eq. (7.6) extended to 2D space. Let us look for a static solution by considering a boundary condition which satisfies

$$V[\phi(r \to \infty)] = 0, \quad \phi \to ve^{-i\alpha(\theta)} \ (r \to \infty) \tag{7.19}$$

The field ϕ has to be continuous to satisfy the differential equation, which imposes a constraint

$$\phi(2\pi) = \phi(0) \tag{7.20}$$

which means

$$\alpha(2\pi) - \alpha(0) = 2\pi n, \quad n = 0, \pm 1, \pm 2, \ldots \tag{7.21a}$$

$$\alpha(\theta) = n\theta + \alpha_0 \tag{7.21b}$$

The Hamiltonian at $r \to \infty$ is given by

$$\mathscr{H} = \frac{1}{2}|\nabla\phi|^2 + V(\phi) = \frac{1}{2}|\nabla\phi|^2 = \frac{n^2v^2}{2r^2} \tag{7.22}$$

which means that the total energy

$$E = \int_0^\infty \int_0^{2\pi} rdr\, d\theta\, \mathscr{H} \tag{7.23}$$

diverges logarithmically. In fact, it is known that vortices of liquid helium in the superfluid state are expressed by such a mathematical expression. The energy density per unit length of the vortex has $\ln L$ dependence on the container size L.

In the field theories, L is always taken to infinity, and the soliton solutions in 1D space as we obtained before cannot be directly extended to 2D space. Therefore, let us add a gauge field to the complex scalar field and change the derivative to the covariant derivative and see what happens. The Lagrangian density becomes

$$\begin{aligned} \mathscr{L} &= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + D_\mu\phi(D^\mu\phi)^\dagger - V(\phi) \\ D_\mu\phi &= \partial_\mu\phi + ieA_\mu \end{aligned} \tag{7.24}$$

This gives the well-known Ginzburg–Landau free energy for a superconductor in 2D space. Let us set $\alpha(\theta) = n\theta$ and choose a gauge (temporal gauge) field that satisfies the condition at $r = \infty$:

$$A^0 = 0, \qquad \mathbf{A} = \frac{1}{e}\nabla(n\theta) = \frac{n}{er}\mathbf{e}_\theta \tag{7.25}$$

The gauge field is a total derivative at $r = \infty$, which can be made to vanish by a suitable gauge transformation. Conversely, it is a field that can be obtained from vacuum by a gauge transformation. Such a field is referred to as a *pure gauge*. The pure gauge satisfies $F_{\mu\nu} = 0$. The above gauge field satisfies the equation

$$D_\theta\phi = \frac{1}{r}\frac{\partial\phi}{\partial\theta} + ieA_\theta\phi = 0, \quad D_r\phi = 0 \tag{7.26}$$

For the static solution, $\mathscr{H} = -\mathscr{L}$ is satisfied, and referring to Eq. (7.22), one sees that $\mathscr{H} \to 0$ as $r \to \infty$ and the total energy can be finite. In other words, there is a possibility that the soliton solution exists.

Let us reconsider the case as a 3D problem which is uniform in the z direction. Then a magnetic flux can exist because we can obtain the total flux in a large enough circle ($r = R$) surrounding the origin ($r = 0$) by

$$\Phi = \int_{r\leq R} B_z dS = \oint \mathbf{A}\cdot d\mathbf{s} = \int_0^{2\pi} A_\theta r d\theta = \frac{2\pi n}{e} \neq 0 \tag{7.27}$$

which means that in the neighborhood of $r \simeq 0$

$$B = B_z = \frac{1}{r}\frac{d}{dr}\left[rA_\theta(r)\right] \neq 0 \tag{7.28}$$

We now see that the effect of adding the gauge field is to create a soliton equipped with magnetic flux. Under this circumstance, the gauge field A_θ cannot be a pure gauge in the neighborhood including $r = 0$. Equation (7.25) holds at $r \to \infty$ but it is singular at $r = 0$. The gauge field given in Eq. (7.25) does not satisfy the continuity condition. Equation (7.26) also means that ϕ has to vanish at $r = 0$. Consequently, in order to satisfy physical conditions, we need to consider that the gauge field A_θ is multiplied by a continuous function which vanishes as $O(r^2)$ at $r \simeq 0$ and reaches 1 asymptotically as $r \to \infty$ (see Figure 7.5a).

As $\phi = 0$ is not a vacuum, there exists a region where the soliton has finite energy density. According to Eq. (7.27), the soliton's magnetic flux is quantized. Despite its 3D structure with an infinitely long magnetic vortex, if it has a finite energy density per unit length and if the magnetic field is uniform in the z direction, the above phenomenon can be interpreted as a 2D soliton solution. Apart from the existence of the complex scalar field, this is the same solenoidal field as the Aharonov–Bohm effect which we discussed in Section 18.3 of [1]. The potential $V(\phi)$ expresses the spontaneously broken scalar field which can be identified with the wave function of condensed Cooper pairs in a superconductor. It is known that, in a superconductor of the second kind, the magnetic flux can penetrate the superconductor, which is referred to as the *Abrikosov's vortex* and is quantized just like Eq. (7.27) claims.

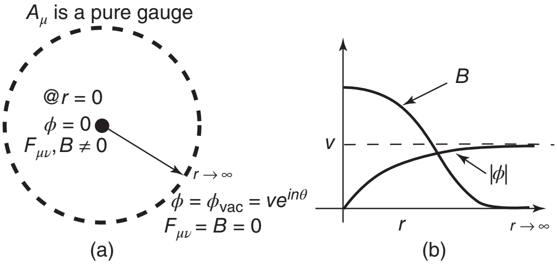


Figure 7.5 Two-dimensional spatial configuration of the magnetic flux soliton B and the spontaneously broken scalar field ϕ. The latter is an order parameter (wave function of the condensed Cooper pairs) inside the superconductor. The magnetic flux invasion depth is about $\delta \sim 1/ev$.

The exact analytical solution to the Lagrangian density in Eq. (7.24) is not known but an approximate solution that satisfies the boundary condition

$$\phi = \chi(r)e^{in\theta}, \quad \chi(r \to 0) \to 0, \qquad \chi(r \to \infty) \to v \tag{7.29}$$

is known and is given by [624]

$$A_\theta = -\frac{n}{er} - \frac{c}{e}K_1(evr) \xrightarrow{r\to\infty} -\frac{n}{er} - \frac{c}{e}\left(\frac{\pi}{2evr}\right)^{1/2} e^{-evr} + \cdots \tag{7.30a}$$

$$B_z = c\chi(r)K_0(evr) \xrightarrow{r\to\infty} \frac{c}{e}\left(\frac{\pi v}{2er}\right)^{1/2} e^{-evr} + \cdots \tag{7.30b}$$

$$\chi(r) \overset{r\to\infty}{\simeq} v(1 - e^{-\sqrt{-m^2}r}), \qquad -m^2 = 2\lambda v^2 \tag{7.30c}$$

K_0, K_1 are modified Bessel functions. A qualitative form of ϕ and B are given in Figure 7.5b. The penetration depth δ, which is the depth that the magnetic flux can penetrate into the superconductor, is given by $\delta \sim 1/ev$.

7.1.3
Winding Number

The vacuum configuration we discussed in the (1 + 2)-dimensional space has a form $\phi = v\exp(-i\alpha)$ and seems to be transferable to each other by a continuous transformation. Then, why is the vortex in 2D space stable? The reason is exactly the same as why the kink is stable, and is based on the topological structure of the vacuum.

a: *U*(1) The configuration for ϕ in Eq. (7.19) expresses a representation of the group $U(1)$ and is a projection from physical vacuum to the representation space of the group. The region of a space to which the vacuum condition is imposed is the boundary of a 2D sphere (the radius $r \to \infty$), that is, a circle (denoted as S^1). If one uses the angular variable θ to specify a position in spacetime, the continuity of ϕ requires identification of $\theta = 2\pi$ with $\theta = 0$. As the representation space also constitutes a circle (S^1), the projection is written as

$$S^1(\text{spacetime}) \longrightarrow S^1(U(1)) \tag{7.31}$$

The group variable $\alpha(\theta)$ is expressed, for instance, like $\alpha(\theta) = n\theta + \beta(\theta)$. The function $\beta(\theta)$ is a continuous function of θ satisfying $\beta(2\pi) = \beta(0)$, and hence can be erased by a suitable continuous transformation. However, two $\alpha(\theta)$'s with differing n cannot be made equal by a continuous transformation. The reason can be pictorially illustrated as in Figure 7.6. Topological difference of two circles with different winding number may become clear if one imagines the elastic ribbon model in Figure 7.4 with two edges identified.

Two different projections with differing numbers of winding cannot be made identical by a continuous transformation. We say that they belong to different homotopy classes. Within the same homotopy class, the projection is a continuous function, but it cannot be transferred by a continuous transformation to a different homotopy class. In this case, there are multiple ϕ's satisfying the same vacuum

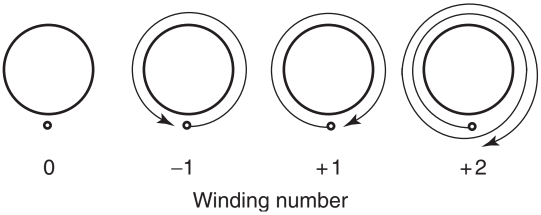


Figure 7.6 Homotopy classification of a projection from S^1 in spacetime to S^1 in the representation space of the group $U(1)$. It is specified by the winding number n. Two topologies with different winding numbers n cannot be deformed to each other by a continuous transformation.

condition but distinguished by an integer which cannot be made identical by a continuous transformation. In mathematical language, it is written as

$$\pi_1(S^1) = Z \tag{7.32}$$

which means that the mapping of S^1 (sub-fix 1 of π) to S^1 is multiply connected specified by an infinite number of integers. Z denotes the ensemble of integer numbers. The integer n is referred to as the *winding number* or *Pontryagin charge.* Using representation variables, it is generically expressed as

$$n = \frac{1}{2\pi}\int_0^{2\pi} d\alpha = \frac{1}{V_U}\int iUdU^{-1} = \frac{1}{V_U}\int dx\, iU\partial_x U^{-1} \tag{7.33}$$

where UdU^{-1} is an operation to extract the group variable, and V_U is the volume of the representation space. At this point, we consider S^1 in spacetime not as a circle but as a line that is obtained by cutting the circle at one point and stretching to infinite length. If both ends of the line ($x = -\infty$ and $x = +\infty$) are identified as the same point, that is, if we regard the phase α as a function of x, rewrite the transformation U as a function of x, and require $U(-\infty) = U(+\infty)$, they are identical as a topology. The expression $U\partial_x U^{-1}$ of the last equality in Eq. (7.33) is nothing but the Jacobian of the transformation of the group variable α to x. We give the following transformation as such an example.

$$U\ :\ \nu(x) = \exp\left[\frac{i\pi x}{\sqrt{x^2+\lambda^2}}\right] \tag{7.34}$$

where λ is an arbitrary constant. It is easily shown that substitution of $U = \nu(x)$ in Eq. (7.33) gives $n = 1$ and that substitution of $U = [\nu(x)]^m$ gives $n = m$.

b: *SU*(2) For the sake of later discussions, we introduce the winding number for $SU(2)$. The $SU(2)$ transformation is expressed generically as

$$U = \exp\left(i\boldsymbol{\alpha}\cdot\frac{\boldsymbol{\tau}}{2}\right) = a + i\boldsymbol{\tau}\cdot\mathbf{b} \tag{7.35}$$

As $a^2 + \mathbf{b}^2 = 1$, this represents the surface of a supersphere in 4D space and can be expressed using three angle variables. From topological arguments, it is expressed as the projection S^3(spacetime)→ $S^3[SU(2)]$, which can also be classified to homotopy classes by a topological number referred to as the *Chern–Simons charge (or number).*[1]

$$Q_{\mathrm{CS}} = \frac{1}{24\pi^2} \oint_{S^3} dS\, \epsilon^{\mu\nu\rho\sigma} n_\mu \mathrm{Tr}[(U\partial_\nu U^{-1})(U\partial_\rho U^{-1})(U\partial_\sigma U^{-1})] \tag{7.37}$$

which takes an integer number. n_μ is a unit vector pointing outward normal to a supersphere embedded in a 4D space.

Using the same argument as used for the $U(1)$, we can use three variables $\mathbf{r} = (x_1, x_2, x_3)$ instead of the three angles and consider S^3 as the whole of the 3D space. Note, however, that all the points at infinity should be identified as the same point. An example of such a representation is given by

$$v(\mathbf{r}) = \exp\left[i\pi \frac{\boldsymbol{\tau}\cdot\mathbf{r}}{\sqrt{r^2+\lambda^2}}\right] \tag{7.38}$$

Here, $U = v^m$ makes the winding number m.

7.1.4
Spacetime Where the Soliton Lives

Derrick's theorem: The conditions for the soliton to exist were discussed for the 1D and 2D space. Derrick's theorem states that the space where the soliton can exist is limited to the following cases only [625]:

1D: Scalar fields (kinks) only
2D: Scalar field + gauge field (magnetic flux in the superconductor)
3D: Scalar field + gauge field (magnetic monopole)
4D: Pure gauge field (instanton).

Proof: We limit our discussions to static solutions in D-dimensional space and simplify arguments by adopting temporal gauge ($A_a^0(x) = 0$, where the suffix "a" denotes internal degrees of freedom). In this case, the Hamiltonian is given by simply changing the sign of the Lagrangian.

$$H = \int \mathscr{H}(\phi, A_\mu) d^D x \tag{7.39a}$$

$$\mathscr{H} = (D_i\phi)(D^i\phi) + \frac{1}{4}F_{a\,ij}F_a^{ij} + V(\phi) \equiv K_{\mathrm{S}} + K_{\mathrm{A}} + V \tag{7.39b}$$

where the index runs from 1 to D. The first term expressed in covariant derivatives is the kinetic energy of the scalar field; the second, the kinetic energy of the gauge field; and the third, the potential energy of the scalar field. We assume

---

[1] In mathematical language, it is expressed as

$$\pi_3(SU(2)) = Z \tag{7.36}$$

See Appendix D

that, by choosing the constants appropriately, each term is adjusted to take only positive-definite values. That is,

$$K_S \geq 0, \quad K_A \geq 0, \quad V \geq 0 \tag{7.40}$$

We consider ϕ and A to be the solutions of the equations of motion. The Hamilton principle constrains them at the extremum for small variations of ϕ or A. Now transform $\phi(x) \to \phi(\lambda x)$ ans $A_i(x) \to \lambda A_i(\lambda x)$, and after finishing differentiation in the Hamiltonian density, set $\lambda x \to x$. The resultant Hamiltonian density is written as

$$\mathscr{H} = \lambda^{2-D} K_s + \lambda^{4-D} K_A + \lambda^{-D} V \tag{7.41}$$

As the energy should take an extremum value at $\lambda = 1$, we differentiate $\mathscr{H}$ by λ and set $\lambda = 1$, and obtain

$$(2 - D)K_S + (4 - D)K_A - DV = 0 \tag{7.42}$$

Considering the constraints Eq. (7.40), we see that only $D = 1$ in the absence of the gauge fields ($K_A = 0$) and only $D = 4$ in the absence of the scalar fields ($K_S = V = 0$) satisfy the condition. We also see that both the scalar and gauge fields can coexist for $D = 2$ or 3, and no solutions exist for $D > 4$.

The topological arguments are useful to investigate the structure of the Lagrangian and check whether it satisfies the necessary condition for the existence of the soliton solution. However, to find whether it exists, one has to solve the equation of motion.

7.1.5
Instanton

We saw that the soliton exists for $D = 3$, which turned out to be the magnetic monopole. It is a theoretically interesting problem. But extensive searches for the monopole have so far been unsucessful. Considering the very powerful inflation scenario in cosmology, which was devised originally to explain the nonexistence of the monopole, we skip discussion of the monopole and concentrate on the soliton in $D = 4$ dimension space. The soliton is localized in time as well as in space, and was given the name ''instanton'' by t'Hooft [626]. Polyakov [627], the other discoverer of the instanton, called it a ''pseudo-particle''.

Degenerate Vacua in *SU*(2) Gauge Theory The instanton is the name given to the pure gauge soliton in 4D Euclidean space. One does not need spontaneously broken vacua for its existence. However, before going into the details of the instanton, we need to discuss the existence of degenerate vacua in 3D physical space in the presence of $SU(2)$ pure gauge fields. In order to discuss the vacuum configuration in the static 3D space, it is convenient to choose the temporal gauge in which the time component of the gauge vanishes ($A_a^0 = 0$). Using bold-faced variables to denote vectors in the isospin space, the gauge transformation to realize it can be obtained by solving the equation

$$\boldsymbol{A}^0(\mathbf{r}) \equiv A_a^0 t_a = U\boldsymbol{A}'^0(\mathbf{r})U^{-1} - (i/g)U\partial^0 U^{-1} = 0 \tag{7.43}$$

where $\mathbf{r}$ denotes the position vector and $t_a = \tau_a/2$ is the $SU(2)$ generators. But even after such a constraint, there still remains a time-independent gauge freedom $\Omega(\mathbf{r})$ which satisfies the condition $\partial^0\Omega(\mathbf{r}) = 0$. Considering that the vacua are the configurations where the field strength[2] satisfies $F_{a\,\mu\nu} = 0$ at $r \to \infty$, the pure gauge that can be obtained from vacua by gauge transformations

$$\boldsymbol{A}_i(\mathbf{r}) = \frac{\tau_a}{2} A_{ai}(\mathbf{r}) = -\frac{i}{g}\Omega(\mathbf{r})\partial_i\Omega^{-1}(\mathbf{r}) \tag{7.44}$$

is static and does not depend on time. Imposing the condition that the gauge fields vanish at infinity, that is

$$\Omega(\mathbf{r}) \to 1 \quad (r \to \infty) \tag{7.45}$$

we can identify all the points at infinity. Therefore, topologically, the 3D space (R^3) with the constraint Eq. (7.45) becomes equivalent to S^3. We already saw that the $SU(2)$ group manifold is S^3, too. Regarding Eq. (7.44) as a mapping from $S^3(\text{spacetime}) \to S^3(SU(2))$, the vacua can be classified to homotopy classes indexed by an integer n defined by Eq. (7.37). In this case, Ω is an ensemble of Ω_n, that is, $\Omega(\mathbf{r}) = \{\Omega_n(\mathbf{r})\}$. Accordingly, the gauge potential is also indexed by n. The corresponding pure gauge potential is given by

$$\boldsymbol{A}_{n,i}(\mathbf{r}) = -\frac{i}{g}\Omega_n\partial_i\Omega_n^{-1} \tag{7.46}$$

An example of such transformations that give $n = 1$ is given in Eq. (7.38).[3] Then $\Omega_n = \nu^n$. It is possible to change $\boldsymbol{A}_n \to \boldsymbol{A}_{n+1}$ by a gauge transformation and it is easily inferred that it is given by ν. Actual calculations show that

$$\boldsymbol{A}_{n+1,i} = -\frac{i}{g}\nu^{n+1}\partial_i\nu^{-n-1} = -\frac{i}{g}\nu\nu^n\partial_i(\nu^{-n}\nu^{-1}) = \nu\boldsymbol{A}_{n,i}\nu^{-1} - \frac{i}{g}\nu\partial_i\nu^{-1} \tag{7.47}$$

Therefore, ν is an operator to change the vacuum with the winding number n to $n+1$. That is,

$$\nu|n\rangle = |n+1\rangle \tag{7.48}$$

In summary, if non-Abelian $SU(2)$ gauge fields exist, there exist degenerate vacua in 3D real space distinguished by the winding number n.

Existence of the Soliton in 4D Space Derrick's theorem tells us that the soliton solution can exist for the pure non-Abelian gauge fields in 4D Euclidean space and that there is no other space where the static soliton can exist. It has also been proved that the existence condition of the soliton is exhausted by $SU(2)$: that is, the soliton solution in any non-Abelian gauge field exists as many times as the number of $SU(2)$ subgroups in the group.

---

[2] Here we use the word "field strength" to denote $F_{\mu\nu}$ to distinguish it from the "field" which is specifically used to denote A_μ.

[3] Another example is $\Omega(\mathbf{r}) = \dfrac{r^2 - \lambda^2 + 2i\lambda\boldsymbol{\tau}\cdot\mathbf{r}}{r^2 + \lambda^2}$

We remind the reader that the existence of the non-Abelian gauge fields in Euclidean 4D space means the existence of the soliton without need of spontaneously broken vacuum. The fact makes the instanton problem unavoidable in the discussions of QCD. The soliton in Euclidean 4D space means a soliton in imaginary time in the Minkowsky space. We defer physical interpretation of the soliton until later, and discuss its mathematical formulation first.

Field theories in the Euclidean space are formulated similarly to those in the Minkowsky space, except that the spacetime variables $x^\mu = (x^1, x^2, x^3, x^4 = ix^0)$ are used. For instance, the action S_E can be obtained from that in the Minkowsky space action S by replacing $x^0 \to -ix^4$, $A^0 \to -iA^4$ and treating x^4 as a real number.

$$S_E = -\frac{1}{2}\int d^4x\,\mathrm{Tr}[\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu}] = -\frac{1}{4}\int d^4x\, F_{a\,\mu\nu}F_a^{\mu\nu} \tag{7.49a}$$

$$\boldsymbol{F}_{\mu\nu} = F_{a\,\mu\nu}t_a, \quad \boldsymbol{A}_\mu = A_{a\,\mu}t_a \tag{7.49b}$$

$$F_a = \partial_\mu A_{a\,\nu} - \partial_\nu A_{a\,\mu} - gf_{abc}A_{b\,\mu}A_{c\,\nu} \quad ^{4)} \tag{7.49c}$$

where f_{abc} is the structure constant of the group in question. Notice, however, that the metric is given by $g^{\mu\nu} = g_{\mu\nu} = (-1, -1, -1, -1)$ and there is no distinction between contravariant and covariant tensors although we retain their upper or lower indices formally. The length in the Euclidean space is expressed as

$$s \equiv \sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2} \tag{7.51}$$

If the gauge fields behave asymptotically as

$$\boldsymbol{A}_\mu \to -\frac{i}{g}U\partial_\mu U^{-1} + O(s^{-2}), \quad s \to \infty \tag{7.52a}$$

then the field strength behaves like

$$\boldsymbol{F}_{\mu\nu} \to 0 + O(s^{-3}), \quad s \to \infty \tag{7.52b}$$

and the Euclidean action S_E becomes finite and satisfies the boundary conditions for the soliton. We now define a quantity

$$\nu \equiv \frac{g^2}{16\pi^2}\int d^4x\,\mathrm{Tr}\left[\tilde{\boldsymbol{F}}_{\mu\nu}\boldsymbol{F}^{\mu\nu}\right] \tag{7.53a}$$

$$\tilde{\boldsymbol{F}}_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}\boldsymbol{F}^{\rho\sigma} \tag{7.53b}$$

---

4) Note that, in the matrix form, the field strength is expressed as

$$\boldsymbol{F}_{\mu\nu} = \partial_\mu\boldsymbol{A}_\nu - \partial_\nu\boldsymbol{A}_\mu + ig[\boldsymbol{A}_\mu, \boldsymbol{A}_\nu] \tag{7.50}$$

and is not gauge-invariant by itself. The action Eq. (7.49a) is gauge-invariant, of course.

The integrand of Eq. (7.53) can be re-expressed in the form of total derivatives

$$\mathrm{Tr}\left[\tilde{\boldsymbol{F}}_{\mu\nu}\boldsymbol{F}^{\mu\nu}\right] = \partial_\mu K^\mu \tag{7.54a}$$

$$K^\mu = \epsilon^{\mu\nu\rho\sigma}\mathrm{Tr}\left[\boldsymbol{A}_\nu\boldsymbol{F}_{\rho\sigma} - \frac{2}{3}ig\boldsymbol{A}_\nu\boldsymbol{A}_\rho\boldsymbol{A}_\sigma\right] \tag{7.54b}$$

Consequently, the variable ν can be rewritten as

$$\nu = \frac{g^2}{16\pi^2}\int d^4x\,\partial_\mu K^\mu = \frac{g^2}{16\pi^2}\int_{S^3} dSn_\mu K^\mu \tag{7.55}$$

If the gauge field satisfies the boundary conditions Eq. (7.52), the first term in Eq. (7.54b) does not contribute and ν coincides with the Chern-Simons number given by Eq. (7.37).[5] Notice that the winding number defined by Eq. (7.53) represents the homotopy classes of the projection from S^3 of the surface of the 4D Euclidean sphere to $S^3(SU(2))$ and is different from that given in the previous section which is the winding number n $[S^3 \to S^3(SU(2))]$ of the projection from 3D space volume to $SU(2)$ manifold.

The soliton solution (instanton) to satisfy the above boundary conditions had been found. An example is given below. Let us first define a gauge transformation U by

$$U(x) = \frac{x_4 + i\boldsymbol{\tau}\cdot\mathbf{r}}{s}, \quad U(x)\partial_\mu U^{-1}(x) = -i\frac{\boldsymbol{\Omega}_{\mu\nu}x^\nu}{s^2} \tag{7.56a}$$

where $\boldsymbol{\Omega}_{\mu\nu}$ is an antisymmetric tensor given by

$$\boldsymbol{\Omega}_{\mu\nu} = \begin{cases} \epsilon^{\mu\nu k}\tau_k & (\mu,\nu = 1 \sim 3) \\ \delta^{\mu k}\tau_k & (\nu = 4) \end{cases} \tag{7.56b}$$

Then the following fields and the field strengths satisfy the conditions for the soliton.

$$\boldsymbol{A}_\mu(x) = -\frac{i}{g}\frac{y^2}{y^2+\lambda^2}U(y)\partial_\mu U^{-1}(y) = -\frac{1}{g}\frac{\boldsymbol{\Omega}_{\mu\nu}y^\nu}{y^2+\lambda^2} \tag{7.57a}$$

$$\boldsymbol{F}_{\mu\nu} = \frac{2}{g}\frac{\lambda^2\boldsymbol{\Omega}_{\mu\nu}}{(y^2+\lambda^2)^2} \tag{7.57b}$$

where $y^\mu = x^\mu - a^\mu$. a^μ and λ are arbitrary constants that represent the position and size of the instanton. This instanton solution satisfies the boundary condition Eq. (7.52) and possesses the following properties which can be proved by performing calculations explicitly:

- It is self-conjugate, that is,

$$\tilde{\boldsymbol{F}}_{\mu\nu} = \boldsymbol{F}_{\mu\nu} \tag{7.58}$$

- $F_{\mu\nu}$ satisfies the non-Abelian field equation of motion.

To prove the statement, one uses the identity $D_\mu\tilde{\boldsymbol{F}}^{\mu\nu} = 0$, which represents sourceless non-Abelian field equations.

---

[5] For the Abelian gauge fields, A_ν and A_ρ commute and give $\nu = 0$.

- It has the winding number $\nu = 1$.
- The action for $\nu = 1$ is given by

$$S_E = -\frac{1}{2}\int d^4x \,\mathrm{Tr}\big[\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu}\big] = \frac{8\pi^2}{g^2} \tag{7.59}$$

- The anti-instanton with the winding number $\nu = -1$ can be obtained by

$$U \to U^\dagger, \qquad \Omega_{\mu\nu} \to \begin{cases} \overline{\Omega}_{\mu\nu} & = \quad \Omega_{\mu\nu}\ (\mu,\, \nu = 1 \sim 3), \\ \overline{\Omega}_{4\nu} & = -\Omega_{4\nu} \end{cases} \tag{7.60}$$

- The instanton for $\nu = m$ can be obtained by changing $U \to U_m = (U)^m$.

Proof: Using polar coordinate variables

$$\begin{gathered} x^\mu = s(\cos\omega, \boldsymbol{n}\sin\omega), \qquad \boldsymbol{n} = (\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta) \\ 0 \le \omega < 2\pi, \quad 0 \le \theta < \pi, \quad 0 \le \phi < 2\pi \end{gathered} \tag{7.61}$$

The transformation function can be rewritten as

$$U = \frac{x_4 + i\boldsymbol{\tau}\cdot\mathbf{r}}{s} = \cos\omega + i\boldsymbol{n}\cdot\boldsymbol{\tau}\sin\omega = \exp(i\omega\boldsymbol{n}\cdot\boldsymbol{\tau}) \tag{7.62}$$

Then $U \to (U)^m$ is equivalent to $\omega \to m\omega$. For x^μ going around S^3 once, the corresponding field goes m times around the rotation axis, which gives the winding number m. ■

Physical Interpretation of the Instanton To understand the physical meaning of the instanton, we deform the boundary at $s \to \infty$ to a circular cylinder (S^c) as described in Figure 7.7.

As the expression Eq. (7.53) for ν is gauge-invariant, we adopt the temporal gauge in which $\boldsymbol{A}_4 = \overline{\boldsymbol{A}}_4 = 0$. Then referring to Eq. (7.55), contributions of the integrand at the side of the cylinder vanish. The top and bottom surface of the cylinder are the 3D volume at $x_4 \to \pm\infty$, with its periphery corresponding to $r \to \infty$. They have precisely the degenerate vacuum configuration that we discussed in Section 7.1.5 and can be assumed to have the winding numbers m and n, respectively. A solution to U to realize $\overline{\boldsymbol{A}}_4 = 0$ can be obtained by substituting $\boldsymbol{A}_4$ in Eq. (7.57) into Eq. (7.43) and is given by the following formulas:

$$U(\mathbf{r}, x_4) = \exp\big[i(\boldsymbol{\tau}\cdot\hat{\mathbf{r}})F(r, x_4)\big], \quad \hat{\mathbf{r}} = \mathbf{r}/r, \quad r = |\mathbf{r}| \tag{7.63a}$$

$$F(r, x_4) = \frac{r}{\sqrt{r^2 + \lambda^2}}\left[\left(n + \frac{1}{2}\right)\pi + \tan^{-1}\frac{x_4}{\sqrt{r^2 + \lambda^2}}\right] \tag{7.63b}$$

The transformation U agrees with v^{n+1} at $x_4 \to \infty$ and v^n at $x_4 \to -\infty$.

$$\begin{gathered} U(\mathbf{r}, -\infty) = v^n(r), \qquad U(\mathbf{r}, +\infty) = v^{n+1}(\mathbf{r}) \\ v(\mathbf{r}) = \exp\left[i\pi\frac{\boldsymbol{\tau}\cdot\mathbf{r}}{\sqrt{r^2 + \lambda^2}}\right] \end{gathered} \tag{7.64}$$

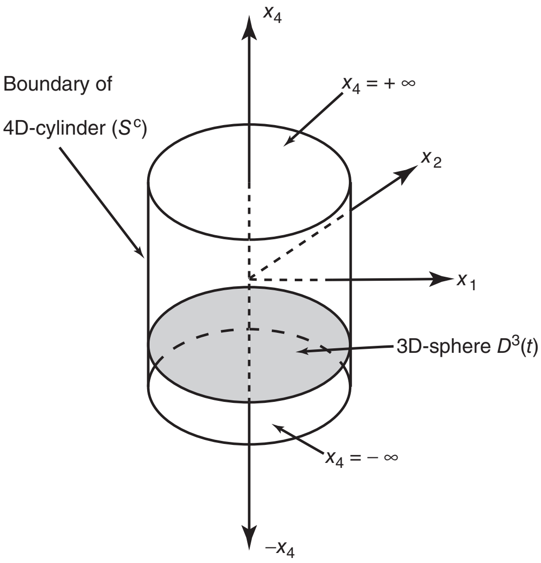


Figure 7.7 The 4D boundary surface (3D volume) modified to a cylinder. The upper surface corresponds to $x_4 = \infty$ and the lower surface to $x_4 = -\infty$. The winding number ν of the projection ($S^3 \to S^3(SU(2))$) defines the topological charge of the instanton. The winding numbers m and n define projections from 3D vacua at ($x_4 = +\infty$) and ($x_4 = -\infty$) to $SU(2)$ configuration ($D^3 \to S^3(SU(2))$). They are related by $\nu = m - n$.

It can also be shown that $\overline{A}_{n,i}$ coincides with $A_{n,i}$ given in Eq. (7.47). That is

$$\begin{aligned} \overline{\boldsymbol{A}}_4(x_4) &= 0 \qquad \text{for all } x_4 \\ \overline{\boldsymbol{A}}_i(x_4 = \infty) &= -\frac{i}{g}v^{n+1}\partial_i v^{-(n+1)} \\ \overline{\boldsymbol{A}}_i(x_4 = -\infty) &= -\frac{i}{g}v^n\partial_i v^{-n} \end{aligned} \tag{7.65}$$

The winding number for this case is given by

$$\begin{aligned} \nu &= \frac{g^2}{16\pi^2}\int d^4x\,\mathrm{Tr}\left[\tilde{\boldsymbol{F}}_{\mu\nu}\boldsymbol{F}^{\mu\nu}\right] \\ &= \frac{g^2}{16\pi^2}\int d^3x\;K^4\Big|_{x_4=-\infty}^{x_4=+\infty} = (n+1) - n = 1 \end{aligned} \tag{7.66}$$

In general, the winding number ν of the instanton is given by the difference of n and m, which is the winding number of the 3D volume in real space at $x_4 = \pm\infty$.

$$\nu = m - n \tag{7.67}$$

In summary, the instanton has the effect of changing a vacuum with winding number n at $x_4 = -\infty$ to that with m at $x_4 = +\infty$, and the difference is given by ν. In other words, the existence of the instanton means transitions of a vacuum to another with a different winding number.

Imaginary Time In classical mechanics, a particle with energy E is bounced back when it collides with a potential whose height is larger than E. The particle cannot penetrate the potential barrier if its peak height exceeds the energy the particle carries. In quantum mechanics, however, the tunneling effect exists with a finite probability. The particles are advancing waves having wave number $p = \sqrt{2m(E - V)}$ and they have position dependence expressed as $\exp(ipx)$. At the place where $V > E$, the wave number is imaginary ($p = i\sqrt{2m(V - E)}$), which represents an attenuating wave. According to semiclassical WKB formalism commonly used in quantum mechanics, the transition amplitude to overcome the potential barrier is given by

$$\exp\left[-\frac{1}{\hbar}\int_a^b \left[2m(V - E)\right]^{\frac{1}{2}} dx\right] = \exp(-S_E) \tag{7.68}$$

Let us evaluate the action in the path from a to b for $E > V$. Normalizing to make the total energy to vanish, we have

$$\hbar S = \int_{t_a}^{t_b} L dt = \int (K - V) dt = \int 2K dt = \int p\frac{dx}{dt} dt = \int_a^b p\, dx \tag{7.69}$$

which differs from the action S_E in Eq. (7.68) only in the sign of $E - V$. On the other hand, the equation of motion is given by

$$m\frac{d^2x}{dt^2} = -\frac{\partial V}{\partial x} \tag{7.70}$$

which tells us that using the imaginary time $x_4 = it$ is equivalent to using the inverted potential. Therefore, the action S_E defined by Eq. (7.68) can be considered as the action in imaginary time. In the field theories, it is represented by the Euclidean action. In other words, the path integral of a Lagrangian in the Euclidean space

$$\exp[-S_E] = \exp\left[-\int \mathscr{L}\right] \tag{7.71}$$

gives the transition amplitude for the path. The reason why we use the action in imaginary time instead of using outcomes of the WKB method is that it can easily be generalized in field theories. In quantized field theories, the path integral method is conventionally used to calculate the transition amplitudes. In the path integral formalism, one needs to integrate over all possible paths. However, contributions of those that are widely apart from the classical path, the ones that minimize the action, are generally small, and a semiclassical calculation generally gives a good approximation. The instanton action was already derived in Eq. (7.59).

$$S_E = -\frac{1}{2}\int d^4x \operatorname{Tr}\left[\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu}\right] = \frac{8\pi^2}{g^2} \tag{7.72}$$

Therefore, the transition probability to a vacuum with a different winding number is given approximately by

$$e^{-\frac{8\pi^2}{g^2}} = e^{-\frac{2\pi}{\alpha_s}} \tag{7.73}$$

In QCD, it is not a small number. The instanton describes the states localized in time as well as in space. But it is not a soliton in the Minkowsky space, and hence is not a particle as we know it. We conclude that, if the instanton exists, there are paths to vacua with different winding numbers and the possibility of transitions from one vacuum to another exists. Because of the potential barriers, the probability per instanton with $\nu = 1$ is of the order $\exp(-8\pi^2/g^2)$.

However, the instanton has aspects very similar to those of ordinary particles. Just like particles, many instantons can exist. If there are many and the average distance between two of them is far greater than their average size, they constitute a dilute gas. If there are many and the average distance becomes small and comparable to their own size, they will behave like a liquid. The state of matter filled with such instanton medium (instanton phase) may be closely connected with the spontaneous breakdown of chiral symmetry [620, 628].

7.1.6
θ Vacuum

We have seen that transitions among vacua with different winding numbers are possible, which means that the vacuum is a superposition of many vacua indexed by n. According to discussions in the previous section, $\Omega = \nu$ in Eq. (7.38) changes n by 1 on acting a vacuum.

$$\Omega|n\rangle = |n+1\rangle \tag{7.74}$$

It means that, generally, the vacuum is not invariant under gauge transformations. However, a gauge-invariant vacuum can easily be constructed as follows, which is referred to as the *θ vacuum*:

$$\begin{aligned} |\theta\rangle &= \sum_n e^{-in\theta}|n\rangle \\ \Omega|\theta\rangle &= \sum_n e^{-in\theta}|n+1\rangle = e^{i\theta}|\theta\rangle \end{aligned} \tag{7.75}$$

It gives a different vacuum for a different value of θ, which is orthogonal to the others and constitutes a completely different world. To see the orthogonality of θ vacua, we calculate the expectation value of a gauge-invariant propagator.

$$\langle G\rangle = \langle\theta|T(J_1\cdots J_p)|\theta'\rangle = \sum_{m,n} e^{i(m\theta-n\theta')}\langle m|T(J_1\cdots J_p)|n\rangle \tag{7.76}$$

As J_i is gauge-invariant, $\Omega^{-1}T(J_1\cdots J_p)\Omega = T(J_1\cdots J_p)$ must hold, which means the right-hand side of the above equality depends only on $\nu = m - n$. Consequently,

$$\langle G\rangle = \sum_m e^{im(\theta-\theta')}\sum_\nu e^{i\nu\theta'}F(\nu) = 2\pi\delta(\theta-\theta')\sum_\nu e^{i\nu\theta}F(\nu) \tag{7.77}$$

In other words, a gauge-invariant perturbation cannot change the value of θ. There is no way to determine the value of θ because the θ vacuum does not interact with external worlds. When the θ vacuum exists, a general transition amplitude can be expressed as

$$<\theta_{\text{out}}|\theta_{\text{in}}>=\sum_{m,n} e^{i(m-n)\theta} <m_{\text{out}}|n_{\text{in}}>=\sum_{\nu} e^{i\nu\theta}\left[\sum_{n} <(n+\nu)_{\text{out}}|n_{\text{in}}>\right] \tag{7.78}$$

which is the sum of transitions among vacua with different winding numbers. Furthermore, each amplitude is multiplied by a phase factor $e^{i\nu\theta}$. It is easily verified using the path integral formalism that the effect can be taken into account by adding the following $\mathscr{L}_\theta$ to the Lagrangian density:[6]

$$\mathscr{L}_{\text{eff}} = \mathscr{L} + \mathscr{L}_\theta = \mathscr{L} + \theta\frac{g^2}{16\pi^2}\int d^4x\,\text{Tr}\left[\tilde{\boldsymbol{F}}_{\mu\nu}\boldsymbol{F}^{\mu\nu}\right] \tag{7.80}$$

In summary, we realize that a very complex entity, referred to as the θ vacuum, exists and that an additional parameter θ has crept into the theory.

The θ vacuum is defined by Eq. (7.75), and can exist in general regardless of the value of θ. $\mathscr{L}_\theta$ is an added term to convert the transition amplitudes for $\theta = 0$ to those in θ vacua with finite θ.

7.1.7
Electroweak Vacua

One interesting consequence of the instanton effect is the nonconservation of baryon numbers in the electroweak (EW) phase transition [630, 669]. In the EW interaction, there are global symmetries to preserve the baryon (quark) number and the lepton number. However, the EW theory being chiral, the existence of the chiral anomaly breaks both the baryon and lepton number conservation. The baryon and lepton currents are defined by

$$J_B^\mu = \frac{1}{3}\sum_{i=1}^{6}\bar{q}_i\gamma^\mu q_i, \qquad J_l^\mu = \sum_{i=1}^{3}\left[\bar{l}_i\gamma^\mu l_i + \bar{\nu}_{iL}\gamma^\mu\nu_{iL}\right] \tag{7.81}$$

In the EW theory, the isospin gauge current is entirely composed of left-handed currents. The hypercharge current B^μ, which couples to the hypercharge $Y = 2(Q - I_3)$, is also chiral: that is, it couples to left-handed fields through the isospin component. Although the baryon current is a vector current, it couples to

---

[6] Using the path integral formalism (see Chapter 11 of [1]), the transition amplitude $<\theta'|e^{-iHt}|\theta>$ from $|\theta>$ to $|\theta'>$ is expressed as

$$\begin{aligned}
<\theta|e^{-iHt}|\theta'> &= \sum_{m,n} e^{im\theta}e^{-in\theta'} <m|e^{-iHt}|n>_J \\
&= \sum_{m,n} e^{im(\theta-\theta')}e^{i(m-n)\theta'}\int(\mathscr{D}A_\mu)_{m-n}\exp\left[i\int d^4x\left(\mathscr{L}+J_\mu A^\mu\right)\right] \\
&= \delta(\theta'-\theta)\sum_{\nu=m-n} e^{i\nu\theta}\int(\mathscr{D}A_\mu)_\nu\exp\left[i\int d^4x\left(\mathscr{L}+J_\mu A^\mu\right)\right] \\
&= \delta(\theta'-\theta)\sum_{\nu}\int(\mathscr{D}A_\mu)_\nu\exp\left[i\int d^4x\left(\mathscr{L}+\mathscr{L}_\theta+J_\mu A^\mu\right)\right]
\end{aligned} \tag{7.79}$$

the anomaly through its left-handed component. The anomaly will be treated in the next section. The vector current can be decomposed into

$$\overline{\psi}\gamma^{\mu}\psi = \frac{1}{2}\left[\overline{\psi}\gamma^{\mu}(1-\gamma^5)\psi + \overline{\psi}\gamma^{\mu}(1+\gamma^5)\psi\right] = \overline{\psi}_L\gamma^{\mu}\psi_L + \overline{\psi}_R\gamma^{\mu}\psi_R \tag{7.82}$$

and calculating for the left-handed and right-handed fields separately, the baryon and the lepton currents obey (see Eq. (7.95))

$$\begin{aligned} \partial_{\mu}J_B^{\mu} &= \partial_{\mu}J_l^{\mu} = \frac{N_g g_W^2}{16\pi^2}\left[-\mathrm{Tr}\left[\tilde{W}_{\mu\nu}W^{\mu\nu}\right] + \frac{1}{2}\tan^2\theta_W \tilde{B}_{\mu\nu}B^{\mu\nu}\right] \\ g_W &= \frac{e}{\sin\theta_W} \end{aligned} \tag{7.83}$$

where N_g is the number of generations, and $W_{\mu\nu}$ and $B_{\mu\nu}$ are the field strengths of the isospin and hypercharge gauge fields. As the right-hand side of Eq. (7.83) is a total derivative [see Eq. (7.54a)], there is no contribution from the Abelian field $B_{\mu\nu}$, but the non-Abelian field $W_{\mu\nu}$ part gives a nonzero contribution because of the instanton effect.

$$\Delta B = \int \frac{dQ_B}{dt}dt = \left[-\int J_B^4 d^3x\right]_{x_4=-\infty}^{x_4=+\infty} \overset{(7.66)}{=\!=} N_g\nu \tag{7.84}$$

Equation (7.84) means that, when the baryon moves in a space where the instanton exists, the baryon number is not conserved. As the lepton doublet couples to the anomaly in the same way as the baryon doublets, we have

$$\begin{aligned} \Delta(B+L) &= 2N_g\nu \\ \Delta(B-L) &= 0 \end{aligned} \tag{7.85}$$

The transition probability due to the instanton is of the order [631, 632]

$$\exp\left[-\frac{16\pi^2}{g_W^2}\right] = \exp\left[\frac{-4\pi\sin^2\theta_W}{\alpha}\right] \sim e^{-170} \tag{7.86}$$

which is too small to have any meaning in ordinary physical processes. However, the above transition probability holds for low temperatures ($T \approx 0$). If the temperature is high ($T \gg m_W$), the transition is possible by overriding the potential through thermal excitation. There exists an instanton-like configuration (referred to as the *sphaleron*) at the saddle point of the potential (top of the potential barrier in Figure 7.4) [633–635] to induce transitions among different vacua indexed by the Chern–Simons number. The sphaleron solution in the temporal gauge ($W_0^a = 0$) is given by

$$W_i^a = \frac{2i}{g_W}\left(i\varepsilon_{iaj}\frac{x_j}{r^2}\right)f(\xi) \tag{7.87a}$$

$$\phi = \frac{v}{\sqrt{2}}i\frac{\boldsymbol{\tau}\cdot\mathbf{x}}{r}\begin{bmatrix}0\\1\end{bmatrix}h(\xi) \tag{7.87b}$$

$$f(0) = h(0) = 0, \quad f(\infty) = h(\infty) = 1 \tag{7.87c}$$

where $r^2 = |\mathbf{x}|^2$, $\xi = r/r_0 = rg_W v$. One sees that the sphaleron solution connects a vacuum state ($|\phi|^2 = v^2/2$) at $r \gg v^{-1}$ and the top of the Higgs potential ($\phi^a = 0$) at $r \ll v^{-1}$. Note that the transition is not quantum tunneling but classical thermal excitations. Unlike the instanton, the sphaleron is an unstable configuration; in fact, the name "sphaleron" means "easy to roll down" in Greek. The transition rate is given approximately by [636]

$$P \sim \exp(-M_{\text{sph}}/T), \quad M_{\text{sph}} = 8\pi \frac{m_W}{g_W^2} \approx 5.5\,\text{TeV} \tag{7.88}$$

where M_{sph} is the mass of the sphaleron interpreted as an unstable particle. At temperature $T \gtrsim M_{\text{sph}}$, the sphaleron induces large transitions among vacua with different baryon or lepton numbers. Therefore, any baryon or lepton number that existed before the EW phase transition era would be wiped out. Note, however, that one interesting feature of the above effect is to preserve $\Delta(B - L)$ as given by Eq. (7.85). The possible outcome of the sphaleron on the baryon asymmetry in the Universe will be discussed in Section 8.6

7.2
Strong CP Problem

7.2.1
Anomaly

QCD is a gauge theory based on the color $SU(3)$. In addition, it contains several other symmetries that characterize contents of the QCD. For instance, in the limit where both masses of the u and d quarks can be neglected, a global symmetry $U(2)_L \times U(2)_R$ holds.[7] When the mass is turned on, the flavor diagonal vector is still conserved but the flavor nondiagonal vector and the axial current are no longer conserved.

$$\partial_\mu V^\mu(x) = \partial_\mu[\overline{u}(x)\gamma^\mu u(x)] = \partial_\mu[\overline{d}(x)\gamma^\mu d(x)] = 0 \tag{7.89a}$$

$$\partial_\mu V_+^\mu(x) = \partial_\mu[\overline{\psi}\gamma^\mu \tau_+ \psi] = \partial_\mu[\overline{u}(x)\gamma^\mu d(x)] = i(m_u - m_d)\overline{u}(x)d(x) \tag{7.89b}$$

$$\begin{aligned}\partial_\mu A_+^\mu(x) &= \partial_\mu[\overline{\psi}(x)\gamma^\mu\gamma^5\tau_+\psi(x)] = \partial_\mu[\overline{u}(x)\gamma^\mu\gamma^5 d(x)] \\ &= i(m_d + m_u)\overline{u}(x)\gamma^5 d(x)\end{aligned} \tag{7.89c}$$

Equation (7.89a) reflects the quark (baryon) number conservation, and Eq. (7.89b) reflects the isospin symmetry which still holds for $m_d = m_u$ if $m_d, m_u \neq 0$. Those symmetries are respected to a high degree in the strong interaction. However, the axial vector part of the symmetry $U(1)_{R-L} \times SU(2)_{R-L}$[8] is broken simply because of the nonzero mass of the quarks. Generally speaking, the reason for the symmetry breakdown stems from one of the following three reasons:

---

[7] In the limit where the mass of the s quark is neglected, an approximate $SU(3)_{\text{flavor}}$ also holds, and then it becomes $U(3)_L \times U(3)_R$. We limit our discussion to $U(2)_L \times U(2)_R$.

[8] The index $R - L$ comes from $\gamma^\mu\gamma^5 = (1/2)[\gamma^\mu(1 + \gamma^5) - \gamma^\mu(1 - \gamma^5)]$.

1) The symmetry is explicitly broken. In this case, a symmetry-breaking term (very often a perturbation) is added to the Lagrangian. A good example is the quark mass term in QCD generated by the symmetry breakdown of the EW interaction which is an external perturbation to QCD. By introducing different masses for the d and u quarks, the isospin symmetry is broken in QCD, as was illustrated in Eq. (7.89b).
2) The symmetry is spontaneously broken or rather hidden. The Lagrangian has the symmetry but the ground state (vacuum) is unstable. The realized stable vacuum breaks the symmetry and the observed phenomena do not exhibit the original symmetry that the Lagrangian had. The symmetry can be broken spontaneously by introducing a self-interacting scalar field like the Higgs in the EW interaction. It can also be broken by the dynamics immanent in its own interaction. A good example is the Cooper pair in QED. The condensed Cooper pairs play the role of the Higgs field and induce spontaneous gauge symmetry breaking and transfer ensemble of electrons to the superconducting phase. Its analog, the quark/antiquark condensate, induces the Nambu–Goldstone vacuum in QCD, which breaks the global chiral symmetry $SU(2)_{R-L}$ in QCD (see Section 7.1.5 of [2]). The effect is additional to the chiral symmetry breaking due to the explicit mass term stated in (1).
3) The symmetry is broken by quantum anomaly. Here, the Lagrangian has the symmetry, and the equation of motion respects it in classical treatments. Nevertheless, the symmetry is broken once quantum effects are introduced. The $U(1)_{R-L}$ (denoted as $U_A(1)$ taken from axial symmetry) is such an example and the subject of our discussion that follows.

Let us consider the following chiral transformation:

$$q \to q' = e^{-i\alpha\gamma^5} q, \qquad q^\dagger \to q'^\dagger = q^\dagger e^{i\alpha\gamma^5} \tag{7.90}$$

Considering $\gamma^5 q_R = \gamma^5[(1+\gamma^5)/2]q = q_R$, $\gamma^5 q_L = \gamma^5[(1-\gamma^5)/2]q = -q_L$, it is a gauge transformation applied in opposite directions for right- and left-handed quarks. The variation of the Lagrangian by an infinitesimal symmetry transformation is given by

$$\delta\mathscr{L} = \left[\frac{\partial\mathscr{L}}{\partial q} - \partial_\mu \frac{\partial\mathscr{L}}{\partial(\partial_\mu q}\right]\delta q + \partial_\mu\left[\frac{\partial\mathscr{L}}{\partial(\partial_\mu q)}\delta q\right] \tag{7.91}$$

The first term vanishes because of the Euler–Lagrange equation of motion, and the second term gives the symmetry-associated Noether current. Substituting $\delta q = -i\delta\alpha\gamma^5 q$, $\delta q^\dagger = i\delta\alpha q^\dagger\gamma^5$ for the infinitesimal chiral transformation into Eq. (7.91), the second term becomes

$$\delta\alpha\partial_\mu(\overline{q}\gamma^\mu\gamma^5 q) \tag{7.92}$$

If the Lagrangian does not contain mass terms ($\mathscr{L}_{\text{mass}} \sim m\bar{q}q$), the QCD Lagrangian is invariant under the chiral transformation and the conserved Noether current

$$\partial_\mu J_5^\mu = \partial_\mu(\bar{q}\gamma^\mu\gamma^5 q) = 0 \tag{7.93}$$

is realized. The existence of the conserved current means the existence of a conserved charge (referred to as the *axial charge*) defined by

$$Q_5 = \int J_5^0(x) d^3x \tag{7.94}$$

In the quantized field theories, the axial charge Q_5 is the generator of the chiral transformation in Eq. (7.90). As stated above, the chiral invariance does not hold if the quark mass is finite. Therefore, we assume zero mass in the following discussion unless otherwise stated.

Now, the above discussions hold only at the Lagrangian level. Once quantum corrections are introduced, it is known that contributions of triangular diagrams like those shown in Figure 7.8 do not vanish.

It was first discovered in the process $\pi^0 \to \gamma\gamma$ and is known as the *Adler–Bell–Jackiew chiral anomaly* [365, 366]. The chiral anomaly generally appears when one axial and two polar vector currents or three axial vector currents couple to a fermion loop. All the quantum anomalies are known to reduce to the triangular diagram contributions. By calculating the contributions of the anomaly in which the chiral current and two gluon fields couple, the equation of the conserved current is modified to

$$\begin{aligned} \partial_\mu J_5^\mu &= 2N_f \frac{g_s^2}{16\pi^2} \text{Tr}\left[\tilde{\boldsymbol{G}}_{\mu\nu}\boldsymbol{G}^{\mu\nu}\right] \\ N_f &= \text{Number of fermion species} \\ \tilde{G}_{\mu\nu} &= \frac{1}{2}\epsilon_{\mu\nu\rho\sigma}G^{\rho\sigma} \end{aligned} \tag{7.95}$$

where $\boldsymbol{G}^{\mu\nu} = G_a^{\mu\nu} t_a$ is the gluon field strength and t_a are the $SU(3)$ group generators. The right-hand side of Eq. (7.95) is the sum of all fermion loops that can contribute to the triangular anomaly. It is irrelevant whether the fermion is massive or not. Equation (7.95) can also be obtained from the Euler–Lagrange equation if $\mathscr{L}_\theta$, introduced in Eq. (7.80) from discussions of the θ vacuum, is added to the Lagrangian.

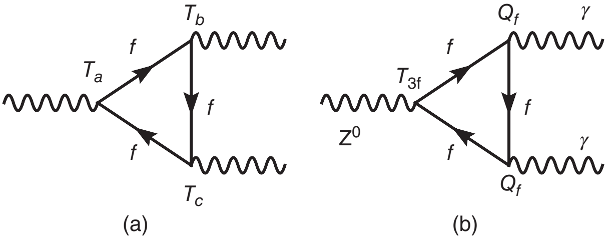


Figure 7.8 Triangular anomaly. Three currents couple to a fermion loop. (a) The current couples to the fermion via the group generator T_a, T_b, T_c. (b) Z and $\gamma\gamma$ couple.

When the anomaly exists, the axial current created as a Noether current of a symmetry is no longer conserved even if the quark mass term is absent. As a result, the Ward–Takahashi or Slavnov–Taylor identity, an essential condition for renormalizability of the theory, is broken. Therefore, if the axial current is a gauge current, it is a serious problem. However, if the theory has a structure such that the sum of all fermion contributions to the loop adds up to zero, the trouble disappears. The standard EW theory has exactly such a structure (see discussions in Section 3.2.3). On the other hand, if the chiral symmetry is global, no serious problem arises. In fact, it has been shown that the anomaly term is necessary to explain the observed decay rate $\Gamma(\pi^0 \to \gamma\gamma)$ [618].

Referring to Eq. (7.91) and (7.92), Eq. (7.95) means that the small variation of the Lagrangian due to an infinitesimal chiral transformation is given by

$$\delta\mathscr{L} = \delta\alpha\partial_\mu J_5^\mu = \delta\alpha\, 2N_f\frac{g_s^2}{16\pi^2}\mathrm{Tr}\left[\tilde{\boldsymbol{G}}_{\mu\nu}\boldsymbol{G}^{\mu\nu}\right] \tag{7.96}$$

Comparing Eq. (7.96) with the extra Lagrangian required by the θ vacuum

$$\mathscr{L}_{\mathrm{eff}} = \mathscr{L}_{\mathrm{QCD}} + \mathscr{L}_\theta = \mathscr{L} + \theta\frac{g_s^2}{16\pi^2}\mathrm{Tr}\left[\tilde{\boldsymbol{G}}_{\mu\nu}\boldsymbol{G}^{\mu\nu}\right] \tag{7.97}$$

we see that the chiral transformation $q \to e^{-i\alpha\gamma^5}q$ changes the value of θ to $\theta + 2\alpha N_f$.

7.2.2
Chiral Transformation and the Mass Term

Now let us consider the effect of the nonzero mass on the chiral transformation of the Lagrangian. Even in the absence of the chiral anomaly, the existence of the quark mass violates conservation of the axial current. Indeed, by calculating the equation of motion, we can derive a relation in the absence of the quantum anomaly.

$$\partial_\mu J_5^\mu = 2im\overline{q}\gamma^5 q \tag{7.98}$$

By the chiral transformation, the mass term in the Lagrangian is modified to

$$-\mathscr{L}_M = m\overline{q}q \quad \to \quad m\overline{q}e^{-i2\alpha\gamma^5}q = m\left(\overline{q}q\cos 2\alpha - i\overline{q}\gamma^5 q\sin 2\alpha\right) \tag{7.99}$$

where the anti-commutativity of γ^0 and γ^5 was used in deriving the above equalities. As the chiral transformation is a gauge transformation for q_L and q_R treated independently, it can also be expressed as

$$q_L \to q_L' = e^{-i\alpha_L}q_L, \qquad q_R \to q_R' = e^{-i\alpha_R}q_R \tag{7.100}$$

Using the following identities and Eq. (7.100),

$$\overline{q}q = \overline{q}_L q_R + \overline{q}_R q_L, \qquad \overline{q}\gamma^5 q = \overline{q}_L q_R - \overline{q}_R q_L \tag{7.101}$$

the mass term after the transformation becomes

$$\overline{q}'q' = \cos(\alpha_R - \alpha_L)(\overline{q}_L q_R + \overline{q}_R q_L) - i\sin(\alpha_R - \alpha_L)(\overline{q}_L q_R - \overline{q}_R q_L) \tag{7.102}$$

Therefore, the two chiral transformations Eq. (7.99) and (7.100) are connected by $\alpha = (\alpha_R - \alpha_L)/2$.

In the Standard Model (SM), the quark mass arises when the EW symmetry is spontaneously broken. Their value is expressed as a product of the coupling constant between the quark and the Higgs and the vacuum expectation value (VEV) v of the Higgs field. The VEV, in general, is not necessarily real. As we learned from the origin of the Cabibbo-Kobayashi–Maskawa matrix, we had to apply different transformations to q_L and q_R to diagonalize the mass matrix (see Section 6.1 in [2]). Consequently, the mass matrix M is not Hermitian in general, which, in turn, means complex mass values. Therefore, writing the mass generically as $m_i = |m_i| e^{i\beta_i}$, the quark mass term is expressed as

$$-\mathscr{L}_M = \sum_i (\overline{q}_{iL} m_i q_{iR} + \overline{q}_{iR} m_i^* q_{iL}) = \sum_i |m_i| \, \overline{q}_i e^{i\beta_i \gamma^5} q_i \tag{7.103}$$

The parameter m_i has to be real to represent a physical mass. It can be achieved by applying chiral transformation individually to each quark. However, every time we apply the chiral transformation, the parameter θ changes to $\theta + 2\beta_i$, and the total change of θ is given by

$$\theta_{\text{EW}} = \sum_i \beta_i = \arg\left(\prod_i m_i\right) = \arg(\det M) \tag{7.104}$$

In summary, the parameter θ that appears as the coefficient of the $\mathscr{L}_\theta$ has two origins: one which we write as $\overline{\theta}$ is required by the θ vacuum, and the other θ_{EW} of EW origin. Therefore, the total θ that appears in the total Lagrangian including both QCD and the EW interactions should be expressed as

$$\theta = \overline{\theta} + \theta_{\text{EW}} \tag{7.105}$$

The above discussion tells us that the parameter θ includes two contributions, one which originates from the multiple vacua of QCD, and the other from the mass phases of the EW interaction. They are considered as two independent effects, and there is no reason for them to compensate each other. We must conclude that $\mathscr{L}_\theta \neq 0$, in general.

The term $\mathscr{L}_\theta$, analog of $\boldsymbol{E} \cdot \boldsymbol{B}$ in QED, violates both P and T and hence CP reversal symmetries. The term is a total derivative, and hence should give no contribution in ordinary dynamics. However, if the instanton exists, the term gives a nonzero value. If such a Lagrangian term exists, it is equivalent to having a CP-violating mass term $\sim im\overline{q}\gamma^5 q$ according to Eq. (7.102). Then, a composite effect combined with the electromagnetic current J^μ_{EM} such as

$$<n| \int d^4x \, \mathscr{L}_\theta J^\mu_{\text{EM}}(0)|n> \sim \theta \int d^4x \; <n|2im(\overline{q}\gamma^5 q) J^\mu_{\text{EM}}(0)|n> \tag{7.106}$$

contributes to the neutron electric dipole moment d_n. According to calculations [637, 638]

$$d_n \sim 4.5 \times 10^{-15} \theta e \cdot \text{cm} \tag{7.107}$$

Comparing with the experimental value [639]

$$d_n < 2.9 \times 10^{-26} e \cdot \text{cm} \tag{7.108}$$

we obtain

$$\theta < 0.7 \times 10^{-11} \tag{7.109}$$

In conclusion, we state that there is no reason to make θ of $\mathscr{L}_\theta$ arbitrarily small, yet its observed value is extremely small if not zero. This is commonly referred to as the *strong CP problem* [618].

7.2.3
U(1) problem

Though the value of θ is undetermined, the existence of the θ vacuum solves at least one pending problem. In QCD, there exists a global flavor chiral symmetry $SU(3)_L \times SU(3)_R \times U(1)_V \times U_A(1)$ in the limit of massless quarks ($m_u = m_d = m_s = 0$). After spontaneous chiral symmetry breaking, the octet and singlet Nambu-Goldstone Boson (NGB) appear as the light scalar mesons with $J^P = 0^-$ ($\pi^\pm, \pi^0, K^\pm, K^0, \overline{K}^0, \eta$ and η'). η and η' are considered mixed mass eigenstates of the $I = 0$ neutral member of the octet η_8 and the singlet η_0. The octet members are considered as the NGB associated with $SU(3)_{R-L}$ and the singlet member as the NGBs associated with $U_A(1)$ chiral symmetry breakdown. With this assumption, and using the current algebra, it follows that

$$m_{\eta'} < \sqrt{3} m_\pi \quad [365, 366, 640] \tag{7.110}$$

However, the observed mass of η' grossly violates the relation ($m_{\eta'} = 975.8$ MeV, $m_{\pi^\pm} = 139.6$ MeV, $m_{\pi^0} = 135.0$ MeV), and hence η' cannot be regarded as the NGB. This is referred to as the *U(1) problem*.

The problem is solved by considering the anomaly that contributes to the chiral current. The argument goes as follows: The existence of the anomaly causes the axial charge to change between two vacua at $t = +\infty$ and $t = -\infty$ due to the instanton effect:

$$\Delta Q_5 \equiv Q_5(t = \infty) - Q_5(t = -\infty) = \int \frac{dQ_5}{dt} dt = \left[\int J_5^0(x) d^3x \right]_{-\infty}^{+\infty} = \nu \neq 0 \tag{7.111}$$

which means nonconservation of the axial current. Then, there is no chiral symmetry associated with $U_A(1)$, and hence no NGB exists to begin with. It follows that η' is not a Goldstone boson and need not respect Eq. (7.110).

So far so good. However, there is some hitch so that we can still argue for the existence of the conserved current. What will happen if we redefine the $U_A(1)$ current by the following relation?

$$\tilde{J}_5^\mu \equiv J_5^\mu - \frac{g_s^2}{16\pi^2} K^\mu \tag{7.112}$$

where J_5^μ is the chiral current and K^μ is defined by Eq. (7.54b). As the K^μ term cancels ν [see Eq. (7.55)], this is a conserved current and the corresponding charge

$$\tilde{Q}_5 = \int \tilde{J}_5^\mu d^3x \tag{7.113}$$

conserves also. Then η' can be considered as the NGB associated with the newly defined chiral current, and the $U(1)$ problem seems to revive. The argument can be refuted as follows:

Let us consider the role of $\tilde{Q}_5$. Referring to Eq. (7.47) and (7.54b), we see that

$$\Omega \left[\frac{g_s^2}{16\pi^2} \int K^0 d^3x \right] \Omega^{-1} = \frac{g_s^2}{16\pi^2} \int K^0 d^3x - 1 \tag{7.114a}$$

$$\therefore \quad \Omega \tilde{Q}_5 \Omega^{-1} = \tilde{Q}_5 + 2N_f \tag{7.114b}$$

Using Eq. (7.75) and (7.114b), we realize that $\tilde{Q}_5$ is an operator to change the value of θ in the θ vacuum.

$$\Omega e^{i\alpha\tilde{Q}_5} |\theta> = \Omega e^{i\alpha\tilde{Q}_5} \Omega^{-1} \Omega |\theta> = e^{i(\theta + 2N_f\alpha)} \left[e^{i\alpha\tilde{Q}_5} |\theta> \right] \tag{7.115}$$

Comparing Eq. (7.115) with Eq. (7.75), we have

$$e^{i\alpha\tilde{Q}_5} |\theta> = |\theta + 2N_f\alpha> \tag{7.116}$$

Thus $\tilde{Q}_5$ has the right property as a symmetry operator. However, by construction, $\tilde{J}_5^\mu$ is not gauge-invariant. One recalls that $\partial_\mu K^\mu$ is gauge-invariant but K^μ is not. The noninvariance is explicitly shown by the relation Eq. (7.114b). Therefore, it cannot affect gauge-invariant observables or transition amplitudes. Consequently, the spontaneous breakdown of the symmetry due to the $\tilde{Q}_5$ operation does not produce a physical NGB, and hence η' cannot be the NGB [641]. It means there is no $U(1)$ problem.

The $U(1)$ problem is considered as a prominent experimental proof for the existence of the θ vacuum.

7.3
Why Do We Need the Axion?

If one of the quarks has vanishing mass, and has freedom to make chiral transformation, the chiral phase parameter can be changed freely and by setting it zero, the θ problem goes away. However, phenomenologically, even the lightest u quark is considered to have a finite running mass from the low-energy phenomena concerning mesons and baryons [642, 643] (see also Section 7.2.5 of [2] for general discussion of the quark masses).

$$m_u(\mu = 1\,\text{GeV}) = 4.6 \pm 0.9\,\text{MeV} \ \ [7] \tag{7.117}$$

We need to find another mechanism to remove the unwanted θ. The most promising scenario to solve the strong CP problem is to require a new symmetry (PQ symmetry after its proposers Peccei and Quinn [618,644,645]) and the existence of a light scalar meson $a(x)$ referred to as the *axion* which is the NGB associated with the symmetry [646, 647]. The axion couples with the gluon through the quantum anomaly, becomes a part of the anomaly, and converts the θ term as follows [see

Eq. (7.131) later]:

$$\mathscr{L}_{a+\theta} = \left(\theta + \frac{a}{f_a}\right) \frac{g_s^2}{16\pi^2} \mathrm{Tr}\left[\tilde{\boldsymbol{G}}_{\mu\nu} \boldsymbol{G}^{\mu\nu}\right] \; {}^{9)} \tag{7.118}$$

As the equation of motion adjusts the position of the potential to a minimum automatically, which is realized at $\overline{\theta} =< \frac{a(x)}{f_a} > +\theta = 0$, the strong CP problem is solved. It means that the CP symmetry in the strong interaction is maintained dynamically.

7.3.1
PQ Symmetry and the Standard Axion

The so-called standard axion, which is the NG boson associated with the original PQ symmetry, was not observed experimentally. Only the possibility of the "invisible axion" remains. However, to understand its role, it is convenient to study the standard axion first and consider the latter as its variant. In the following, we restrict our discussion to u and d quarks and add other quarks when necessary. The original idea was to extend the notion of the chiral charge to fields other than quarks, namely, to the Higgs. Let us consider fermion–Higgs Yukawa interactions

$$\begin{gathered}
\mathscr{L}_m = \frac{\sqrt{2}m_u}{v_2}\overline{\Psi}_L \Phi_2 u_R + \frac{\sqrt{2}m_d}{v_1}\overline{\Psi}_L \Phi_1 d_R + h.c. \quad {}^{10)} \\
\Psi_L = \begin{bmatrix} u_L \\ d_L \end{bmatrix}, \qquad \Phi_1 = \begin{bmatrix} \varphi^+ \\ \varphi^0 \end{bmatrix}, \qquad \Phi_2 = \begin{bmatrix} \varphi'^0 \\ \varphi'^- \end{bmatrix}
\end{gathered} \tag{7.119}$$

In the SM, $\Phi_2 = \Phi_1^c = i\tau_2 \Phi_1^\dagger$. If the scalar fields also carry the chiral charge (denoted as Q_{PQ}), they can be gauge-transformed by the chiral transformation.

$$\begin{gathered}
\Psi_L \to e^{-i\alpha_L}\Psi_L, \quad u_R \to e^{-\alpha_{u_R}} u_R, \quad d_R \to e^{-\alpha_{d_R}} d_R \\
\Phi_{k\,(k=1,2)} \to e^{-i\alpha_k}\Phi_k
\end{gathered} \tag{7.121}$$

If the phases satisfy the relations

$$\alpha_1 = \alpha_L - \alpha_{d_R}, \qquad \alpha_2 = \alpha_L - \alpha_{u_R} \tag{7.122}$$

the Yukawa interaction is invariant under the PQ symmetry operation. Notice that the invariance cannot be achieved in the SM because, for the PQ symmetry to hold, one needs at least two independent Higgs doublets to satisfy Eq. (7.122), that is, two Higgs with different Q_{PQ} charge. In the SM, the Higgs carry the electric charge Q as well as the hypercharge Y. One has to set part of the Higgs carrying the

---

9) Different authors use different definitions for the axion decay constant f_a. It is related to the magnitude v of the VEV that breaks the $U_A(1)$ symmetry by $f_a = <a> /N$, where N is a model-dependent parameter.

10) Notice that, unlike the formalism adopted in the discussion of the supersymmetry, we are back to the conventional four-component Dirac spinor formalism.

$$\overline{\Psi}_L \Phi_1 = \overline{u}_L \varphi^+ + \overline{d}_L \varphi^0 \tag{7.120}$$

PQ charge not to mix with that carrying Y, because after EW symmetry breaking the latter is absorbed by the third component of the neutral Z boson. If it mixes, the Z will exert an additional force not allowed in the SM. Extension of the SM to include two Higgs doublet was discussed earlier (see Chapters 1 and 5). Two charged Higgs $H^{\pm}$, three neutral Higgs h^0, H^0, and A appear. h^0, H^0 are CP-even and carry hypercharge, while A is CP-odd and does not carry the hypercharge. We need a combination of the scalar fields to carry the Q_{PQ} but not the hypercharge, because otherwise the SM would contain an unwanted extra interaction originating from the Q_{PQ} charge. Consequently, we can adopt the configuration of A for the axion which has the form [see Eq. (1.57)]

$$a \sim -\text{Im}(\varphi_1^0)\frac{v_2}{v} + \text{Im}(\varphi_2^0)^c\frac{v_1}{v}, \qquad v = \sqrt{v_1^2 + v_2^2} \tag{7.123}$$

where $v_i = \sqrt{2} < \varphi_i^0 >$ are the VEVs of the Φ_i fields. It is convenient to adopt the form of phase field. Then, independent of the other components, the axion can be built into Φ_i's as

$$\begin{aligned} \Phi_1 &= \frac{1}{\sqrt{2}}v_1\exp\left(\frac{ia}{xv}\right)\begin{bmatrix}0\\1\end{bmatrix}, \qquad \Phi_2 = \frac{1}{\sqrt{2}}v_2\exp\left(\frac{iax}{v}\right)\begin{bmatrix}1\\0\end{bmatrix} \\ x &= \frac{v_1}{v_2} \end{aligned} \tag{7.124}$$

For the adopted form of the axion field, the PQ transformation in Eq. (7.122) changes the field to

$$a \to a + \alpha v \tag{7.125}$$

Assignment of the PQ charge to each field is model-dependent. In the original paper, all the left-handed fermions are assigned with $Q_{\text{PQ}} = 0$ and d_R and u_R are assumed to couple only to Φ_1 and Φ_2, respectively. In this case, we have

$$\alpha_{d_R} = \frac{\alpha}{x}, \qquad \alpha_{u_R} = \alpha x \tag{7.126}$$

The $U_A(1)$ current carrying Q_{PQ} can be calculated to be

$$J^{\mu}_{\text{PQ}} = -v\partial^{\mu}a + x\overline{u}_R\gamma^{\mu}u_R + \left(\frac{1}{x}\right)\overline{d}_R\gamma^{\mu}d_R \tag{7.127}$$

As the chiral current has the color anomaly contribution, it satisfies the relation

$$\partial_{\mu}J^{\mu}_{\text{PQ}} = \left(x + \frac{1}{x}\right)\frac{g_s^2}{16\pi^2}\text{Tr}\left[\tilde{G}_{\mu\nu}G^{\mu\nu}\right] \tag{7.128}$$

To investigate the relation of the axion with $\mathscr{L}_{\theta}$, one only needs to make a transformation to erase the axion formally in Φ_i. By the operation, the fermions are transformed by

$$u_R \to e^{-i\frac{ax}{v}}u_R, \qquad d_R \to e^{-i\frac{a}{xv}}d_R \tag{7.129}$$

As a result, the mass terms acquire a phase

$$\frac{a}{v}\left(x + \frac{1}{x}\right) \equiv \frac{a}{v}N \tag{7.130}$$

but according to discussions leading to Eq. (7.103) and Eq. (7.105), it is equivalent to changing the value of θ to

$$\theta \to \theta + \frac{a}{v} N \tag{7.131}$$

where N is a model-dependent coefficient. The term

$$\left(\theta + \frac{a}{v} N\right) \frac{g_s^2}{16\pi^2} \mathrm{Tr}\left[\tilde{\boldsymbol{G}}_{\mu\nu} \boldsymbol{G}^{\mu\nu}\right] \tag{7.132}$$

provides a potential for the axion. Excitations of a are not from $<a>= 0$ but from the bottom of the potential. Therefore

$$\begin{aligned} a &=<a> + a_{\text{phys}} \\ \theta &+ \frac{<a>}{v} N = 0 \end{aligned} \tag{7.133}$$

a_{phys} represents the physical axion, which is renamed as a in the following. The above equation means that the color anomaly provides an additional potential to the axion and hence the wine bottle potential is tilted, and the axion fields falls into the newly created potential minimum.

Even if $\theta \neq 0$ in the beginning, dynamics of the axion readjusts the vacuum position in such a way as to guarantee the vanishing of θ.

Viewing the relation Eq. (7.129) from the fermion current perspectives, it provides the axion interaction with the fermions. This is because the fermion kinetic energy induces the derivative couplings to the axion. Writing down the resultant Lagrangian that contains the axion field, it is given by [648]

$$\begin{aligned} \mathscr{L}_{\text{axion}} = \frac{1}{2} \partial_\mu a \partial^\mu a &+ \frac{1}{v} \left[\partial_\mu a \sum_i \left(x \bar{u}_{iR} \gamma^\mu u_{iR} + \frac{1}{x} \bar{d}_{iR} \gamma^\mu d_{iR} \right) \right. \\ &\left. + N a \frac{g_s^2}{16\pi^2} \mathrm{Tr}\left[\tilde{G}_{\mu\nu} G^{\mu\nu}\right] \right] \\ &\qquad + \frac{1}{v} \left[\partial_\mu a \frac{1}{x} \sum \bar{l}_{iR} \gamma^\mu l_{iR} + E a \frac{e^2}{32\pi^2} \tilde{B}_{\mu\nu} B^{\mu\nu} \right] \end{aligned} \tag{7.134a}$$

$$B^{\mu\nu} = F_{\text{EM}}^{\mu\nu} - \tan\theta_W F_Z^{\mu\nu} \tag{7.134b}$$

$$N = N_{\text{g}} \left(x + \frac{1}{x} \right), \qquad E = 2N_{\text{g}} \left(\frac{4}{3} x + \frac{1}{3x} + \frac{1}{x} \right) \tag{7.134c}$$

Here, we included the contributions of all the fermions. That is, in the first big bracket $\left[\cdots\right]$, we included all the fermions in the strong interaction and in the second $\left[\cdots\right]$ we included the equivalent of the first bracket that can be obtained by extending arguments to the EW sector. N_{g} denotes the number of generations, $B^{\mu\nu}$ is the $U(1)$ gauge field strength which is a mixture of the electromagnetic field $F_{\text{EM}}^{\mu\nu}$ and the neutral current field $Z(F_Z^{\mu\nu})$, and θ_W is the Weinberg angle. The parameters N and E are model-dependent. The axion acquires mass radiatively through the chiral anomaly term. An effective mass term $-m_a^2 a^2$ is created by the virtual process $a \to \text{GG} \to a$ where GG is the loop process. Therefore, the axion mass is proportional to the expectation value of the coefficient of a in Eq. (7.134a).

Its value can be calculated by using low-energy current algebra or by formulating a low-energy effective Lagrangian. We only quote the result [649].

$$m_a^2 = \frac{m_\pi^2 F_\pi^2}{v^2} N^2 \frac{m_u m_d}{(m_u + m_d)^2} \quad \rightarrow \quad m_a \sim 75 \left(x + \frac{1}{x} \right) \text{keV} \tag{7.135}$$

$F_\pi \sim 93$ MeV is the pion decay constant. The axion a has the same quantum number as π and η, and mixing with them is possible. As the coupling constant of the axion with the fermions is given by $\sim 1/v = (\sqrt{2}G_F)^{1/2}$, it is fundamentally a weak interaction.

The existence of the standard axion is almost completely denied by laboratory experiments and astrophysical arguments. The most solid evidences come from experiments listed in Table 7.1.

In addition, it is supported by various astrophysical arguments discussed in more detail later including, for instance, the cooling rate of the red giants. As long as the scale v_{PQ} of the PQ symmetry breakdown is given by the EW scale $v_{\text{PQ}} \sim 1/(\sqrt{2}G_F)^{1/2}$, there remains no space for the axion to exist. Consequently, if one seeks the origin of vanishing $\mathscr{L}_\theta$ in the PQ symmetry, the value of v_{PQ} has to be made very large, which terrestrial experiments up to now have been unable to detect. Such axions are generically referred to as the *invisible axions* although the name is a bit of a misnomer because specially designed instruments may be able to detect it.

7.3.2
Invisible Axion

There are a variety of models for the invisible axion. In general, it appears as the phase field of a scalar field σ which has the following properties in common:

(1) It carries the PQ charge Q_{PQ} and has a very large VEV.

(2) It is a singlet in terms of $SU(2) \times U(1)$ in the SM. It means it interacts with ordinary particle only through the $U_{\text{PQ}}(1)$ symmetry.

The necessity of the item (2) is to avoid inducing the large-scale v_{PQ} in the SM interaction. Models differ depending on what Q_{PQ} is assigned to ordinary particles.

Table 7.1 Experiments to deny the existence of the standard axion.

Reaction	Theory	Experiment
$K^+ \rightarrow \pi + a$	$\sim 3 \times 10^{-5}(x + 1/x)^2$	$< 3.8 \times 10^{-8}$ [650]
$J/\psi \rightarrow a + \gamma$	$3.7 \pm 0.8 \times 10^{-5} x^2$	$< 1.4 \times 10^{-5}$
$\Upsilon \rightarrow a + \gamma$	$2.0 \pm 0.7 \times 10^{-4}(1/x)^2$	$< 3 \times 10^{-4}$ [651]
$\tau(a \rightarrow 2\gamma)$	$\sim 0.2\,(100\,\text{keV}/m_a)^5$ sec	not detected
$\tau(a \rightarrow e^+e^-)$	$\sim 4 \times 10^{-9}(1\,\text{MeV}/m_a)(1 - 4m_e^2/m_a^2)^{1/2} x^3$ sec	not detected

We list two representative models (KSVZ and DFSZ) below. The names are initials of the authors.

KSVZ axion [652, 653]: The ordinary fermion carries no $Q_{\rm PQ}$ but assumes the existence of a new (probably very heavy) quark X which carries $Q_{\rm PQ}$.

$$-\mathscr{L} = h\overline{X}_L\sigma X_R + h.c., \qquad \sigma = \frac{v_{\rm PQ}}{\sqrt{2}}e^{\frac{ia}{v_{\rm PQ}}} \tag{7.136}$$

Assuming only X_R carries the $Q_{\rm PQ}$, arguments similar to those of the standard axion tell us that the axion can couple to the ordinary particles through the anomaly.

$$\mathscr{L}_{\rm KSVZ} = \frac{a}{v_{\rm PQ}}\left[\frac{g_s^2}{16\pi^2}{\rm Tr}\left[\tilde{\boldsymbol{G}}_{\mu\nu}\boldsymbol{G}^{\mu\nu}\right] + 3(Q_{\rm XEM})^2\frac{e^2}{16\pi^2}\tilde{F}_{{\rm EM}\,\mu\nu}F^{\mu\nu}_{\rm EM}\right] \tag{7.137}$$

where $Q_{\rm XEM}$ is the electric charge of the X quark and $F_{{\rm EM}\,\mu\nu}$ is the electromagnetic field strength. The interaction is almost identical to the standard axion except for the replacement $v \to v_{\rm PQ}$. The KSVZ axion does not couple to the electron at the tree level and is referred to as the *hadronic axion*.

DFSZ axion [654, 655]: The ordinary fermions carry $Q_{\rm PQ}$. Consequently, the model requires two Higgs fields just like the standard axion. However, the ordinary fermions cannot couple directly to σ but can couple through the Higgs potential. The field σ couples with the ordinary Higgs through quartic coupling.

$$\mathscr{L}_{\sigma H} = \kappa\Phi_1^T C(\sigma^\dagger)^2\Phi_2 + h.c. \tag{7.138}$$

In the limit $v_{\rm PQ} \gg v$, the three Higgs fields assume the following form.

$$\sigma = \frac{v_{\rm PQ}}{\sqrt{2}}\exp\left(\frac{ia}{v_{\rm PQ}}\right) \tag{7.139a}$$

$$\Phi_1 = \frac{v_1}{\sqrt{2}}\exp\left(\frac{iX_1 a}{v_{\rm PQ}}\right)\begin{bmatrix}0\\1\end{bmatrix}, \qquad \Phi_2 = \frac{v_2}{\sqrt{2}}\exp\left(\frac{iX_2 a}{v_{\rm PQ}}\right)\begin{bmatrix}1\\0\end{bmatrix} \tag{7.139b}$$

$$X_1 = \frac{2v_2^2}{v^2}, \qquad X_2 = \frac{2v_1^2}{v^2} \tag{7.139c}$$

The σ interaction in Eq. (7.138) is invariant under the transformation $a \to a + \alpha v_{\rm PQ}$. Couplings to other particles have the same form as the standard axion except for the replacements $v \to v_{\rm PQ}$, $x \to X_2$, and $1/x \to X_1$. Only the interaction strength is smaller by a factor $(2v_1v_2/vv_{\rm PQ})$. The DFSZ axion has a tree-level interaction with the lepton.

Finally, let us give values of the mass and coupling strength, which will be useful in discussing phenomenology. They appear in the effective Lagrangian, which is the result of current algebra or phenomenological Lagrangian including higher order corrections [7, 656, 657].

$$\begin{aligned} m_a &= \frac{m_\pi F_\pi}{f_a}\frac{\sqrt{z}}{1+z} = 6.3\,\mu\text{eV}\left(\frac{10^{12}\,\text{GeV}}{f_a}\right)\text{[11]} \\ f_a &= \frac{v_{\text{PQ}}}{N}, \quad F_\pi = 93\,\text{MeV}, \quad z = \frac{m_u}{m_d} = 0.56 \pm 0.042\ \ [658] \end{aligned} \tag{7.140}$$

The interaction of the axion with the photon γ and the fermion f can be summarized as follows [659, 660]:

$$\mathscr{L}_{a\gamma\gamma} = -\frac{1}{4}g_{a\gamma\gamma}aF_{\mu\nu}\tilde{F}^{\mu\nu} = -g_{a\gamma\gamma}a\boldsymbol{E}\cdot\boldsymbol{B} \tag{7.141a}$$

$$g_{a\gamma\gamma} = \frac{e^2}{8\pi^2 f_a}\left(\frac{E}{N} - \frac{2}{3}\frac{4+z}{1+z}\right) = \frac{\alpha}{2\pi}\left(\frac{E}{N} - \frac{2}{3}\frac{4+z}{1+z}\right)\frac{1+z}{\sqrt{z}}\frac{m_a}{m_\pi F_\pi} \tag{7.141b}$$

$$N = \sum_j Q_{\text{PQ}j}, \qquad E = 2\sum_j Q_{\text{PQ}j}Q_j^2 N_{cj} \tag{7.141c}$$

N and E are model-dependent coefficients that appear in the color anomaly and the electromagnetic anomaly. $E/N = 8/3$ for the grand unified theories (DFSZ model included) and $E/N = 0$ for the KSVZ model (for $Q_{XEM} = 0$). $Q_{\text{PQ}j}$ is the PQ charge of the jth fermion, Q_j is the electric charge in units of the proton charge, and N_{cj} is the color degrees of freedom (three for quarks and one for leptons). The reason why the color coefficients N appear in the photon coupling is because the axion mixes with π^0 and η. The two-photon decay width is

$$\begin{aligned} \Gamma(a \to \gamma\gamma) &= \frac{g_{a\gamma\gamma}^2 m_a^3}{64\pi} = \frac{\alpha^2}{256\pi^3}\frac{m_a^5}{m_\pi^2 F_\pi^2}\left[\left(\frac{E}{N} - \frac{2}{3}\frac{4+z}{1+z}\right)\frac{1+z}{\sqrt{z}}\right]^2 \\ &= 1.1\times 10^{-24}s^{-1}\left(\frac{m_a}{\text{eV}}\right)^5 \end{aligned} \tag{7.141d}$$

where the second expression uses Eq. (7.141b) with $z = 0.56$ and $E/N = 0$. As $g_{a\gamma\gamma} \propto 1/f_a \propto m_a$, the constraint on the axion models draws a belt for the allowed region on the $m_a - g_{a\gamma\gamma}$ plane (see Figure 7.12 and Figure 7.13). One may consider more general ALPs in which the two parameters are made independent.

The interaction with fermions is given by

$$\mathscr{L}_{af} = -g_{aff}\overline{\psi}_f i\gamma^5\psi_f a \tag{7.141e}$$

$$g_{aff} = \frac{C_f m_f}{f_a}\text{[12]}, \qquad \alpha_{aff} = \frac{g_{aff}^2}{4\pi} \tag{7.141f}$$

The dimensionless combination $g_{aff} = C_f m_f/f_a$ plays the role of a Yukawa coupling constant and α_{aff} of a "fine-structure constant". m_f is the fermion mass and C_f is a model-dependent parameter (see Table 7.2).

---

[11] The mass–decay constant relation $m_a f_a \approx m_\pi F_\pi$ is an essential feature of the axion. There are many models that require similar scalar fields. In phenomenological treatments, the axion is often generalized to general axion-like particles (ALPs) in which the parameters m_a and f_a are set independent.

[12] For the equivalent differential coupling, the Lagrangian is given by $\mathscr{L}_{aff} = -(C_f/2f_a)\partial_\mu a\overline{\psi}_f\gamma^\mu\gamma^5\psi$.

Table 7.2 Coupling constants in DFSZ and KSVZ models

	DFSZ	KSVZ
$g_\gamma = (E/N - 1.92)/2$	0.37	-0.96
C_e	$(1/3)\cos^2\beta$	0
C_p	$-0.10 - 0.45\cos^2\beta$	-0.39
C_n	$-0.18 + 0.39\cos^2\beta$	$+0.04$

$\tan\beta = v_2/v_1$, $g_\gamma = g_{a\gamma\gamma}(\pi f_a/\alpha)$

Note that g_γ *and* $C_f(f = e, p, n)$ are dimensionless, while $g_{a\gamma\gamma}$ has the dimension m^{-1}

In Table 7.2, we list the numbers that are often used for phenomenological analysis. From the table, one sees that the KSVZ axion does not couple to the electron directly and that the coupling strength to the photon is about 10 times larger compared with the DFSZ model.

7.4
Constraints on Invisible Axions

The axion, if it exists, not only solves an important theoretical problem but also plays an important role in astrophysics [661, 662]. For some range of parameter spaces, it is also an important dark matter candidate. The outstanding characteristics of the axion are its small mass and the weak interaction very similar to those of the neutrino. Because of this, the role of the axion in the stellar evolution can be derived directly from that of the neutrino with minor modifications. However, from the cosmological point of view, it differs from the neutrino in one important aspect. It is a cold dark matter[13] in contrast to the hot dark matter (HDM) (moving at relativistic speed at the time of decoupling) of the neutrino. It is a result of the axion being at rest and behaving as a condensed Bose–Einstein object.

7.4.1
Coolant of the Stellar Evolution

If the axions are made during the stellar evolutionary processes, they carry their energy directly away to the outside of the star. That is, the axion acts as a very efficient coolant, accelerating the stellar evolution and reducing its life. The axion interactions that act as coolants include the following:

(1) Compton scattering : $\gamma + e \to e + a$
(2) Axion Bremsstrahlung : $e + Z \to e + Z + a$
(3) Primakoff effect : $\gamma + Z \to Z + a.$

Imposing the condition that the axion reaction does not disturb the known stellar evolution processes, one can constrain the range of values of the mass or the coupling constants. By choosing different stellar processes, one can constrain the coupling strength of the axion to those of the electron, the nucleon, and the photon separately. In the following, we introduce some of the astrophysical constraints on the coupling constants [656, 657].

---

[13] Axions, if produced thermally, behave almost identically to the neutrino as a hot dark matter. But the possibility of the thermal axion is excluded except for a very few special cases. We concentrate our discussions on the axion as a cold dark matter.

Horizontal branch stars in globular clusters: An average star, comparable to the Sun in size, is in the so-called main sequence. A main-sequence star generates energy by burning hydrogen fuel (i.e., through $4\mathrm{p} + 2\mathrm{e}^- \rightarrow {}^4\mathrm{He} + 2\nu_e$). When the hydrogen fuel is exhausted, it transitions from a main sequence star to a red giant where helium burning plays an important role. After repeating several helium flashes in the beginning of the transition phase, it passes through the so-called horizontal branch (HB) in the Heltzsprung–Russel (color– magnitude; see also Section 11.1.2) diagram and advances to AGB (Figure 7.9).

In the horizontal branch, the helium burning at the core and the hydrogen burning at the surface of the helium core are subtly balanced, keeping the total

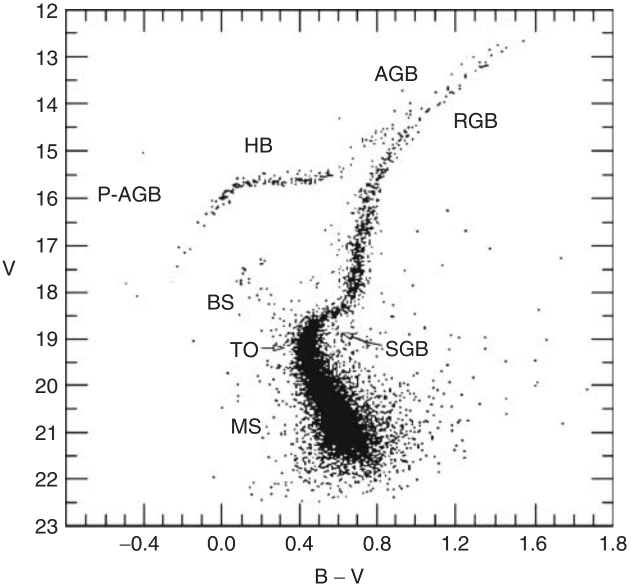


Figure 7.9 Color–magnitude diagram for the globular cluster M3, based on 10,637 stars [663]. Shown vertically is the brightness in the visual (V) band, and horizontally the difference between B (blue) and V brightness, that is, a measure of the color and thus surface temperature, where blue (hot) stars lie toward the upper left. The classification for the evolutionary phases is as follows: MS (main sequence): core hydrogen burning. BS (blue stragglers). TO (main-sequence turnoff): central hydrogen is exhausted. SGB (subgiant branch): hydrogen burning in a thick shell. RGB (red-giant branch): hydrogen burning in a thin shell with a growing core until helium ignites. HB (horizontal branch): helium burning in the core and hydrogen burning in a shell. AGB (asymptotic giant branch): helium and hydrogen shell burning. P-AGB (post-asymptotic giant branch): final evolution from the AGB to the white-dwarf stage (Reproduced with permission of [662].) See also Figure 11.4 for the stellar evolution process.

luminosity at a constant value while decreasing the surface temperature (hence its name). As the core temperature is about 10^8 K ($\sim$10 keV), axions with $m_a \lesssim 10$ keV contribute to the cooling process. Here, they are produced mainly through the Primakoff process. The energy release by the stars in this stage is $\sim$80 erg $g^{-1}s^{-1}$, compared to Primakoff axion losses of g_{10}^2 30 erg $g^{-1}s^{-1}$ where $g_{10} \equiv g_{a\gamma\gamma}/10^{-10}$ GeV^{-1}. During the process, the neutrino cooling is very effective, and the existence of the axion sharply disturbs (accelerates) the process. It reduces the HB lifetime by about $80/(80 + 30 g_{10}^2)$, as well as the number of HB stars. In particular, stars in a globular cluster, which are considered to be born at the same time, are distributed in the various phases of the stellar evolution depending on their size. The HB lifetime can be measured relative to the standard stellar evolution rate (i.e., the number of RGBs, which are much brighter and hence not much affected by the Primakoff effect) and provides especially useful information on the axion–photon coupling.

$$g_{a\gamma\gamma} < 1 \times 10^{-10}\ \mathrm{GeV}^{-1} \quad [657, 662] \tag{7.142}$$

Helium ignition and white dwarfs: If the axion couples directly to electrons, the dominant emission processes are $\gamma + e \to e^- + a$ and $e^- + Z \to Z + e^- + a$. The enhanced energy losses would delay helium ignition. This implies that the core grows more massive before the helium ignition and pushes the RGB to brighter stars. Brightness of the brightest stars in the RGB branch of globular clusters is a sensitive probe of the extra axion emission.

Moreover, the Bremsstrahlung is efficient in white dwarfs where the Primakoff and Compton processes are suppressed by the large plasma frequency. The observed luminosity functions of the white dwarfs agree with the SM. The obtained constraint gives a similar bound to that obtained from the RGB.

$$\alpha_{aee} < 1.3 \times 10^{-27} \quad [664–666] \tag{7.143}$$

This is equivalent to $g_{aee} < 1.3 \times 10^{-13}$ or in the DFSZ model to

$$f_a > 1.3 \times 10^9\ \mathrm{GeV} \quad \text{and} \quad m_a < 4.5\ \mathrm{meV} \quad \text{for } \cos^2\beta = \frac{1}{2} \tag{7.144}$$

SN1987A: Similar constraints from the observation of neutrino bursts from the supernova SN1987A can be obtained [629, 667, 668]. But this time we treat confinement and liberation of the axion in the core of the neutron star. That is, constraints on the axion–hadron couplings can be obtained. The contributing process is axion Bremsstrahlung by nucleons $N + N \to N + N + a$. The axion, if it exists, amplifies the role of the neutrino and reduces the burst time. It happens when the coupling is small and gives a lower limit to the coupling strength. On the other hand, if the coupling is strong, more axions are captured in the core, which reduces the cooling effect. Therefore, the coupling strength cannot be too large, either. The obtained constraints are

$$3 \times 10^{-10} < g_{aNN} < 3 \times 10^{-7} \tag{7.145}$$

If the coupling strength becomes even larger, the axion itself as well as the neutrino will be observed. From observations, the following limit is obtained [669]:

$$10^{-6} \lesssim g_{aNN} \lesssim 10^{-3} \tag{7.146}$$

The limits on the coupling strength can be converted to those on f_a or m_a. Using $C_p = -0.4$ and $C_n = 0$, the limit Eq. (7.145) means [7]

$$f_a \gtrsim 4 \times 10^8\,\mathrm{GeV} \quad \text{and} \quad m_a \lesssim 16\,\mathrm{meV} \tag{7.147}$$

There is a small gap between Eq. (7.145) and Eq. (7.146) from the SN1987A observation alone, but it is filled by observations of globular clusters.

We summarize the limits obtained from astrophysical arguments in Figure 7.10 [7].

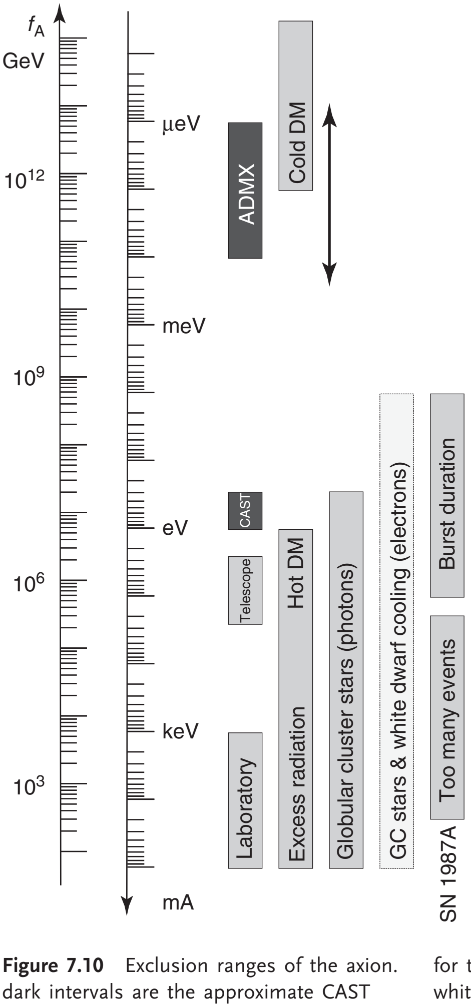


Figure 7.10 Exclusion ranges of the axion. dark intervals are the approximate CAST and ADMX search ranges (to be explained later in the text). The ''Laboratory'' bar is a rough representation of the exclusion range for the standard axion. The ''GC stars and white-dwarf cooling'' range uses the DFSZ model with an axion–electron coupling corresponding to $\cos\beta^2 = 1/2$. (Reproduced with permission of [7].)

7.4.2
Axion as the Dark Matter

Thermal axion: The axion, if it exists, is produced in early Universe. We consider the possibility of the axion remaining to this day as a part of the dark matter. First, we consider a possibility of the axion as a hot dark matter[14]. The process of the thermal axion to become dark matter is very similar to that of the neutrino (see Section 10.3.2). The axion couples to the gluon through the anomaly, and hence it is in thermal equilibrium if $T \gg \Lambda_{\text{QCD}}$. For $T < \Lambda_{\text{QCD}}$, the axion mainly interacts with the pion and, at $T < 30$ MeV, it decouples from the thermal equilibrium. If it is stable enough to remain to this day, the number of remnant axions should be comparable to that of the neutrinos. Consequently, the mass of the remnant axions can be obtained similarly to the neutrino [671, 672].

$$m_a < 0.7 \text{ eV} \tag{7.148}$$

However, the axion can decay to two photons with a decay rate given by Eq. (7.141d). Therefore, axions with $m_a \gtrsim 20$ eV have a shorter lifetime than the age of the Universe ($\sim 4.3 \times 10^{17}$ s). It should have decayed away by now. Under certain circumstances, the photons from the axion decay should be observable as an excess in the photon spectra of the standard cosmology. So far, no excess has been observed and the axion with mass larger than $\gtrsim 1$ eV is excluded (see Figure 7.10 for the excluded range denoted as Hot DM).

Consequently, the possibility of the thermal axions to be a main component of the dark matter can be considered as excluded except in very special circumstances. What is more interesting is the so-called misalignment axion produced via nonthermal processes, which is our main subject in the following [125, 673].

7.4.3
Misalignment axion

The misalignment axions as the dark matter are formed in two steps. The first step occurs at the cosmic temperature $T \sim v_{PQ}$ when the Higgs field σ develops VEV due to spontaneous breakdown of the $U_{\text{PQ}}(1)$ symmetry. The second step occurs at $T \lesssim \Lambda_{\text{QCD}} \sim 200$ MeV when the axion feels the color anomaly. The effective Lagrangian in the latter stage is given by

$$\mathscr{L} = \left(\theta_{\text{eff}} + \frac{a}{f_a}\right) \frac{g_s^2}{16\pi^2} \text{Tr}\left[\tilde{\boldsymbol{G}}_{\mu\nu} \boldsymbol{G}^{\mu\nu}\right] \tag{7.149}$$

The angle θ_{eff} is given by the VEV acquired in the first stage with additional contribution of the quark mass phase added later after the EW phase transition. Generation of the axion mass occurs at $T \sim \Lambda_{\text{QCD}}$. The mass value develops with time[15] and it is only after $T \ll \Lambda_{\rm QCD}$ that it settles down to an asymptotic value which remains to the present time. In the first step, the field σ takes the form given by Eq. (7.136). As the axion is defined as the phase field of σ, its VEV $<a(x)>/f_a$ can take an arbitrary value in the range between $-\pi$ and $+\pi$. At the second stage, as soon as the axion begins to feel the color anomaly, the wine bottle potential is tilted and the axion falls down to the valley [the value of $<a>$ that satisfies Eq. (7.133)] of the tilted potential. This is the origin of the name ''misalignment axion'' [125, 673].

---

[14] See [670] for details of thermal axions.

The axion, however, does not approach the valley asymptotically. It overruns the bottom of the valley and begins to oscillate. $<a(x)>$ is a classical field, corresponding not to a particle but some collective motion behaving as a field. The potential that the axion feels has been activated by the anomaly that has as many valleys as the number of winding number. One cannot predict which of the n valleys the axion chooses, but here we assume for simplicity that $n = 0$ was chosen. Putting $\Theta(x) \equiv <a(x)>/f_a$, the Lagrangian to describe motion of the axion may be approximated by

$$\mathscr{L} \simeq f_a^2 \left[\frac{1}{2}\dot{\Theta}^2 - \frac{1}{2}m_a^2\Theta^2\right] \tag{7.150}$$

as is always the case in the neighborhood of the potential minimum. Equation (7.150) produces the familiar equation of motion for a harmonic oscillator.

In the expanding universe, the scale R of the space is also changing as a function of time. The cosmic expansion effect can be taken into account by changing the action to $\int \mathscr{L} R(t)^3 d^4x$. The resultant equation of motion for Θ is given by

$$\begin{aligned} &\ddot{\Theta} + 3H(t)\dot{\Theta} + m_a^2(t)\Theta = 0 \\ &H(t) = \frac{\dot{R}}{R} \end{aligned} \tag{7.151}$$

where $H(t)$ is the Hubble expansion rate at time t. At this stage, we recapitulate some basic facts in cosmology just to refresh the readers' memory[16]. Denoting the cosmic temperature as T, the cosmic energy density ρ as a function of the cosmic scale factor R decreases $\sim R^{-4}$, $R \sim T^{-1} \sim t^{1/2}$ in the radiation-dominated early Universe. When the Universe becomes matter-dominated, ρ behaves as $\sim R^{-3}$, $R \sim T^{-1} \sim t^{2/3}$. It happens shortly before the photon decoupling. At temperature $T \gg \Lambda_{\rm QCD}$, the axion is massless and the solution to Eq. (7.151) is Θ = constant. Furthermore, it is a real constant in the sense that it does not depend on the space position either. This is because, even if the field had position dependence, its Fourier component was completely red-shifted from the time of $U_{\rm PQ}(1)$ symmetry breakdown ($T \sim v_{\rm PQ}$) to the QCD era ($T \sim \Lambda_{\rm QCD}$) by the factor $\sim v_{\rm PQ}/\Lambda_{\rm QCD}$. The axion field, then, is static and ultracold, and is in a state of Bose condensate. At the time when the axion begins to feel the color anomaly, the mass m_a begins to grow. At the time when $m_a(t) \approx H(t)$, Θ begins to oscillate (damped oscillation). If m_a grows slowly (adiabatic approximation: $dm_a/dt \ll m_a^2$), the solution can be obtained by replacing $\dot{\Theta}^2$ with its average over the period $<\dot{\Theta}>^2 = \lambda(t)$, $\lambda(t) = (\dot{\Theta}^2 + m_a^2\Theta^2)/2$. Multiplying Eq. (7.151) by $\dot{\Theta}$ and rewriting it in terms of λ, we obtain

---

[15] The field produces particles (energy quanta) when it oscillates in the neighborhood of the potential minimum. The mass (squared) is a name given to the coefficient of the quadratic term in the Taylor expansion of the potential around the minimum. As the potential develops as a function of time, so does the mass value.

[16] Some knowledge of cosmology is required to understand contents of the axion as the dark matter. Those who are not familiar with basic cosmology are recommended to read Chapters 8 first.

$$\frac{\dot{\lambda}}{\lambda} + 3H - \frac{\dot{m}_a}{m_a} = 0 \tag{7.152a}$$

$$\therefore \quad \lambda(t) = A\frac{m_a(t)}{R^3} \tag{7.152b}$$

It means that the axion makes a damped oscillation, with the amplitude developing as $\sim R^{-3/2}$ as the Universe expands. Denoting t_{osc} as the time when the oscillation begins, the energy density of the axion is given by

$$\rho_a(t_{\text{osc}}) \sim \frac{1}{2}f_a^2 m_a^2(t_{\text{osc}})\Theta^2(t_{\text{osc}}) = \frac{C}{2}f_a^2 m_a^2(t_{\text{osc}}) \tag{7.153}$$

C is a constant of order 1, which is determined by the initial value of Θ $(0 \le \Theta < 2\pi)$. At $t > t_{\text{osc}}$, the energy density develops as

$$\rho_a(t) \sim f_a^2 \lambda(t) \tag{7.154}$$

After m_a settles down at the value of $T = 0$ at $T \ll \Lambda_{\text{QCD}}$, ρ_a changes as $\rho_a = n_a m_a \sim m_a/R^3$.

To summarize, through the period during which the value of m_a is changing, the number per expanding volume $R^3(t)$ is kept constant. As stated before, the axion behaves like a Bose condensate with zero momentum, and the oscillation is a collective motion of many bodies. At this stage, the Universe is still dominated by radiation. Consequently, the cosmic energy density is decreasing as $\rho \sim R^{-4}$, which means the relative axion density ρ_a/ρ is growing fast. For a large value of f_a, the axion's contribution to the present cosmic density is non-negligible.

As the number per expanding volume is kept invariant, the present $(t = t_0,\ T = T_0)$ density is given by

$$\rho_a(t_0) = \rho_a(t_{\text{osc}})\frac{m_a}{m_a(t_{\text{osc}})}\left(\frac{T_0}{T_{\text{osc}}}\right)^3 \tag{7.155}$$

By looking at the equation of motion, the time t_{osc} can be defined by the relation

$$3H(t_{\text{osc}}) = m_a(t_{\text{osc}}) \tag{7.156}$$

Using the time–temperature relation [Eq (8.60)] in the thermal Universe, we can estimate the temperature at that time by

$$m_a(t_{\text{osc}}) = 3H(t_{\text{osc}}) \sim \frac{T_{\text{osc}}^2}{M_{Pl}} \tag{7.157}$$

where M_{Pl} is the Planck mass. Substituting $m_a \sim m_\pi F_\pi/f_a$ [Eq. (7.140)], $T_{\text{osc}} \sim \Lambda_{\text{QCD}} \sim 200$ MeV, we obtain an approximate value of the present axion density.

$$\rho_a(t_0) = \frac{C}{2}f_a\frac{m_\pi F_\pi T_0^3}{M_{\text{Pl}}\Lambda_{\text{QCD}}} \sim \frac{f_a \Lambda_{\text{QCD}}}{M_{Pl}}T_0^3 \tag{7.158}$$

The requirement that it does not exceed the present cosmic critical density constrains that

$$\begin{aligned} &\rho_a < \rho_c = 3H^2(t_0)/8\pi G_{\text{Newton}} = 11h^2\,\text{keV/cm}^{-3} \qquad (h \simeq 0.72) \\ \rightarrow\quad &f_a < 10^{12}\,\text{GeV} \quad (\text{or } m_a \gtrsim 10^{-5}\text{eV}) \end{aligned} \tag{7.159}$$

A recent detailed estimate gives [661]

$$\Omega_a h^2 \approx 0.7\left(\frac{f_a}{10^{12}\,\text{GeV}}\right)^{\frac{7}{6}}\left(\frac{<\Theta(0)>}{\pi}\right)^2 \approx 0.3\left(\frac{f_a}{10^{12}\,\text{GeV}}\right)^{\frac{7}{6}} \tag{7.160}$$

Here, $\Omega_a = \rho_a/\rho_c$, $<\Theta(0)>$ is the average initial value. Considering the recent observed value $\Omega_{\text{DM}}h^2 = 0.1187 \pm 0.0017$ by Planck [674], we see that the axion of mass

$$m_a \simeq 6\,\mu\text{eV}\left(\frac{10^{12}\,\text{GeV}}{f_a}\right) \qquad (f_a < 0.5 \times 10^{12}\,\text{GeV}) \tag{7.161}$$

contributes to the cosmic dark matter. Although the axion mass is very small, it is in the state of Bose condensation at the time of matter-radiation equality. Therefore it is a cold dark matter (CDM). The excluded range for the axion mass from the above arguments is denoted as "Cold DM" in Figure 7.10.

The present velocity of the axion can be obtained from the red-shifted momentum of its initial value $p_a \sim m_a$ and gives

$$v_a \sim \frac{p(t_0)}{m_a(t_0)} \sim \frac{m_a(t_{\text{osc}})}{m_a(t_0)}\left(\frac{T_0}{T_{\text{osc}}}\right) \sim 10^{-23}\left(\frac{m_a}{\text{eV}}\right)^{-1} \tag{7.162}$$

which can be considered as completely at rest.

7.5
Laboratory Axion Searches

As a prelude to the experimental searches [675], we summarize the main properties of the axion.

1) The axion, if exists, solves the strong CP problem.
2) It is a pseudo-scalar particle which has $J^P = 0^-$.
3) Its fundamental properties are almost uniquely determined once its VEV f_a is given.
4) Its mass is given by $m_a \sim F_\pi m_\pi/f_a$.
5) The coupling strength with the quark is $\sim m_q/f_a$.
6) The axion couples to the electromagnetic field in the form $g_{a\gamma\gamma}a\boldsymbol{E}\cdot\boldsymbol{B}$ with the strength $g_{a\gamma\gamma} \sim \alpha/(\pi f_a)$.
7) The coupling with the lepton is model-dependent. At least the coupling by the mechanism (6) exists. There exists a model in which the coupling is as strong as that with the quark.
8) For $m_a = 10^{-3} \sim 10^{-6}$ eV, the axion could be a dominant component of the dark mater.

Where does the axion reside? Laboratory searches for the axion denied the existence of the standard axion with mass greater than $O(\text{keV})$. Those with small mass but larger than ~ 1 eV contradict with known facts of stellar evolutions. Nevertheless, we will discuss axion searches in the mass range below ~O(keV) because we do not want to depend too much on astrophysical arguments. If the mass is larger than 0.1 eV, thermally produced axions dominate, but below it the misalignment axions become dominant. There is a possibility that the axions are mass-produced at the core of the Sun. If the mass is below $\sim 10^{-3}$ eV, they may exist as dark matter. In particular, for the mass below $\sim 10^{-5}$ eV, it can be the dominant component of the dark matter. The outstanding characteristic of the axion is that, despite its near-vanishing mass, it is nonrelativistic and is a CDM.

The reason for the axion to be a dark matter candidate comes from the following consideration: It rarely interacts with matter, and the gravitational growth of the mass density fluctuation commences as soon as the transition from radiation-dominated to matter-dominated Universe occurs (see Section 8.7.1). Prior to the cosmic photon decoupling at which point baryon fluctuations start to grow, a considerable amount of fluctuation was formed by the dark matter and rather strong gravitational centers existed in abundance. Baryons are attracted to these gravitational centers, forming structures that develop eventually to galaxies.

As a CDM, the axions are captured by the galaxies and should exist as a halo. From the observed rotational velocity of our galaxy, the energy and number density of the dark matter are estimated to be [676]

$$\begin{aligned} \rho_{\text{Halo}} &\sim 0.45\,\text{GeV cm}^{-3} \\ n_a &\sim 3 \times 10^{13}/\text{cm}^{-3} \left(\frac{\rho_a}{450\text{MeV cm}^{-3}} \right) \left(\frac{10^{-5}\ \text{eV}}{m_a} \right) \end{aligned} \tag{7.163}$$

We discuss several methods to detect the axion in different mass ranges. As for the thermal axions, it is possible to detect their decay photons which are monochromatic. Below 10^{-5} eV, a promising method is to use the Primakoff effect where the axion is converted to a photon in intense radio frequency magnetic fields [677, 678]. Above the mentioned energy range, a method to detect the axion by converting it to X-rays in the magnetic field is useful. It was utilized to detect axions produced in the Sun [679, 680].

Axion Helioscope There is a possibility that the axion is produced abundantly in the Sun by the Primakoff effect. As the core temperature of the Sun is $\sim 2 \times 10^7\ K \sim 1.7$ keV, the helioscope is tuned to detect the axion with energy in a few keV range. The flux on the earth is estimated to be [662]

$$\begin{aligned} \frac{d\Phi_a}{dE} &= g_{10}^2\, 6.0 \times 10^{10} \text{cm}^{-2} s^{-1} \text{keV}^{-1} E^{2.481} e^{-\frac{E}{1.205}} \\ g_{10} &= g_{a\gamma\gamma} \times 10^{10}\,\text{GeV} \end{aligned} \tag{7.164}$$

where E is in keV. The integrated flux is given by

$$\Phi_a = g_{10}^2\, 3.75 \times 10^{11}\ \text{cm}^{-2}\text{s}^{-1}, \tag{7.165}$$

which, in turn, means the solar axion luminosity [681]

$$L_a = g_{10}^2\,1.85 \times 10^{-3} L_\odot, \qquad L_\odot = 3.86 \times 10^{33}\ \text{erg s}^{-1} \tag{7.166}$$

where $L_\odot$ is the solar luminosity. The maximum of the distribution is at 3.0 keV and the average is 4.2 keV with a small dependence on the solar model. As the axion effect modifies the outcome of the standard solar model, we can gain some constraints on the axion properties. Comparison of the sound speed profile diagnosed by helioseismology or the temperature dependence of ^8B neutrino flux constrains

$$g_{a\gamma\gamma} \lesssim 5 \sim 10 \times 10^{-10}\ \text{GeV}^{-1} \tag{7.167}$$

which corresponds to $L_a \lesssim 0.04 - 0.2\,L_\odot$. More restrictive constraints can be obtained by direct observations of the solar axion flux. When the axion passes through a strong magnetic field, X-rays of $2 - 20$ keV and average ~ 4 keV are emitted. As the axion interacts with the electromagnetic field in the form $\sim g_{a\gamma\gamma} a\mathbf{E}\cdot\mathbf{B}$, the magnetic component in the direction of the photon polarization contributes. Consequently, by making the magnetic field perpendicular to the solar direction, one can detect the converted X-rays as the axions pass through it. The conversion probability P_γ is proportional to the Fourier component of the magnetic field [679]. When the magnetic field is approximately uniform

$$P_\gamma(E, L) = \left| \frac{g_{a\gamma\gamma}}{2} \int_0^L dz\, B(z) e^{iz\Delta q} \right|^2 = \left[\frac{g_{a\gamma\gamma} BL}{2} \right]^2 \left[\frac{\sin \frac{\Delta q L}{2}}{\frac{\Delta q L}{2}} \right]^2 \tag{7.168}$$

where Δq is the longitudinal momentum transfer from the axion to the photon and L is the length of the magnetic field. Denoting the photon momentum as $p_\gamma = \omega = E_a$, the energy conservation leads to

$$\Delta q = p_{\gamma z} - p_{az} \simeq \omega - \sqrt{E_a^2 - m_a^2} \simeq \frac{m_a^2}{2\omega} \tag{7.169}$$

When $\Delta q L \gg 1$, the conversion rate is small because of oscillation. Consequently, the effective length useful for observation is given by the condition $\Delta q L_{\text{eff}} \lesssim 1$. For $\omega = 4.2$ keV,

$$L_{\text{eff}} \sim \frac{2\pi}{\Delta q} = \frac{2\pi}{\left(\frac{m_a^2}{2\omega}\right)} = 1.0\,m \left(\frac{0.1\ \text{eV}}{m_a} \right)^2 \tag{7.170}$$

which constrains the detection range to below $m_a \lesssim 0.1$ eV. However, by using a dielectric material to fill the path of the axion, its sensitivity can be extended to larger masses [680]. The wavenumber, that is, momentum in this case, is changed from $p_\gamma \to np_\gamma$, and hence one has to use the following formula instead of Eq. (7.169):

$$\Delta q = p_{\gamma z} \left(1 - \frac{\omega_p^2}{\omega^2} - i\frac{\Gamma}{2\omega} \right)^{\frac{1}{2}} - (E_a^2 - m_a^2)^{\frac{1}{2}} \sim \frac{m_a^2 - \omega_p^2}{2\omega} \tag{7.171}$$

where $\omega_p^2 = 4\pi\alpha n_e/m_e$ is the plasma frequency, and Γ is the absorption rate. Therefore, by choosing a suitable gas that satisfies $\omega_p \sim m_a$, one can keep L_{eff} to a reasonable length. Accordingly, the conversion rate is also modified to

$$P_\gamma = \frac{(g_{a\gamma\gamma} B/2)^2}{|\Delta q|^2 + \Gamma^2/4}\left(1 + e^{-\Gamma L} - 2e^{-\Gamma L/2}\cos\Delta q L\right) \tag{7.172}$$

Early helioscope searches were performed at Brookhaven [682] and Tokyo [683–685]. The first helioscope search that reached the "axion line" (see the band in Figure 7.11) was carried out by the CAST (CERN Axion Solar Telescope) group at CERN. They fabricated a detector having $B = 9.0$ T and length $L = 9.26$ m [681]. The instrument can be adjusted ±8° vertically and observe for 1.5 h during the sunrise and sunset. They also extended the sensitivity region of m_a up to 0.4 eV by filling the space of the magnetic field with pressurized helium [686].

No signal was obtained, and an upper limit was set.

$$g_{a\gamma\gamma} < 8.8\times 10^{-11}\,\text{GeV}^{-1} \quad 95\%\text{CL} \quad m_a < 0.02\,\text{eV} \quad [681, 686, 687] \tag{7.173}$$

The limit is competitive in the mass range $10^{-2} - 10^{-6}$ eV with that obtained from the horizontal branch stars.

Telescope Searches The contribution of the thermal axion to the cosmic energy is small [$\Omega \sim 0.01(m_a/\text{eV})$], but its decay life is short [$\tau(a \to 2\gamma) \sim 10^{25} s\,(m_a/\text{eV})^5$] and the emitted monochromatic photon may be detected. As the axion is virialized (see Appendix A), the photon is not completely monochromatic but Doppler-broadened ($\Delta\lambda/\lambda \sim 10^{-2}$). The predicted intensity at the surface of the earth is comparable to twilight at nights. In order to eliminate the twilight background, one observes the

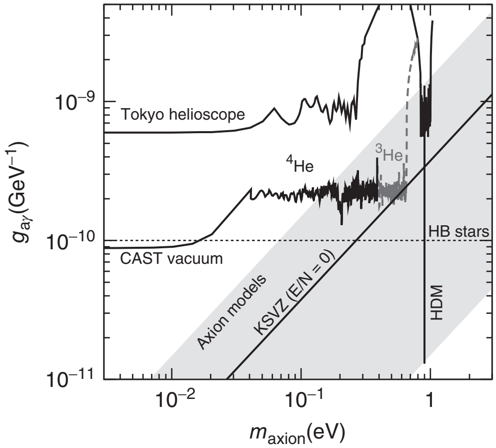


Figure 7.11 Solar exclusion plot for axion-like particles CAST09 [687]. Also shown are constraints from the Tokyo helioscope [683–685]. The dashed line is the sensitivity of the ongoing ^{3}He phase of CAST. The vertical line (HDM) is the hot dark-matter limit [671, 688]. The shaded band represents models with $0.07 < |E/N - 1.92| < 0.7$, and the diagonal solid line corresponds to KSVZ axions. (Reproduced with permission of [7].)

axion trapped by clusters of galaxies. The axion should exhibit the same red shift as the galaxies. By subtracting the light from other than the galaxies, the background is eliminated but the light from the axion remains. A search using the telescope at Kitt Peak National Observatory [689] and a more recent search in two rich Abell clusters [690] were made. No signals were observed. As a result, the axion with mass in the range denoted as ''Telescope'' in Figure 7.10 and 7.14 was excluded.

Microwave Cavity Detectors The microwave detector [691, 692], similar to the solar telescope, relies on the Primakoff effect to detect the axion-converted photon in the resonant cavity, but its sensitivity region is below $m_a \simeq 10^{-5}$ eV. A 5μeV axion at rest would convert to a 1.2-GHz photon. In this method, a crucial point in making a workable detector depends on how far the noise level of the signal amplifier can be lowered. The U.S. experiments focused on improving coherent microwave amplifiers (photons as waves), while the Japanese experiment worked to develop a Rydberg-atom single-quantum detector (photons as particles) [693–695]. As the latter is still in the development stage, we will focus on the U.S. experiments.

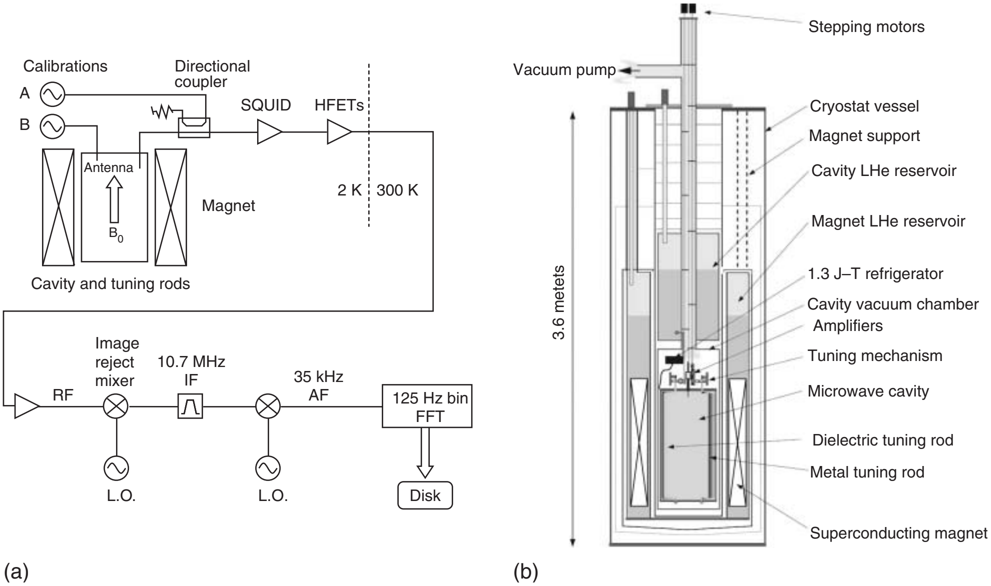

Calibrations
A
B
Directional coupler
SQUID
HFETs
Antenna
B_0
2 K
300 K
Magnet
Cavity and tuning rods
Image reject mixer
10.7 MHz IF
35 kHz AF
RF
125 Hz bin FFT
L.O.
L.O.
Disk
(a)
Stepping motors
Vacuum pump
Cryostat vessel
Magnet support
Cavity LHe reservoir
Magnet LHe reservoir
1.3 J–T refrigerator
Cavity vacuum chamber
Amplifiers
Tuning mechanism
Microwave cavity
Dielectric tuning rod
Metal tuning rod
Superconducting magnet
3.6 metets
(b)

Figure 7.12 Left: Schematic of the ADMX experiment. Photons, created in the cavity by conversion of the axions, are picked up by the antenna and amplified by the SQUID (Superconducting QUantum Interference Devices) and HFETs (heterostructure field-effect transistors). The signal is mixed in two stages, with band-limiting filtering in between, to audio frequencies. The audio spectrum is measured and stored to the disk. The sweep oscillator A provides a reflection measurement to enable adjustment of antenna coupling; the oscillator B, weakly coupled, allows measurement of the cavity resonant frequency. Right: Overview of ADMX hardware including the superconducting magnet and the cavity insert (Reproduced with permission of [696–698].)

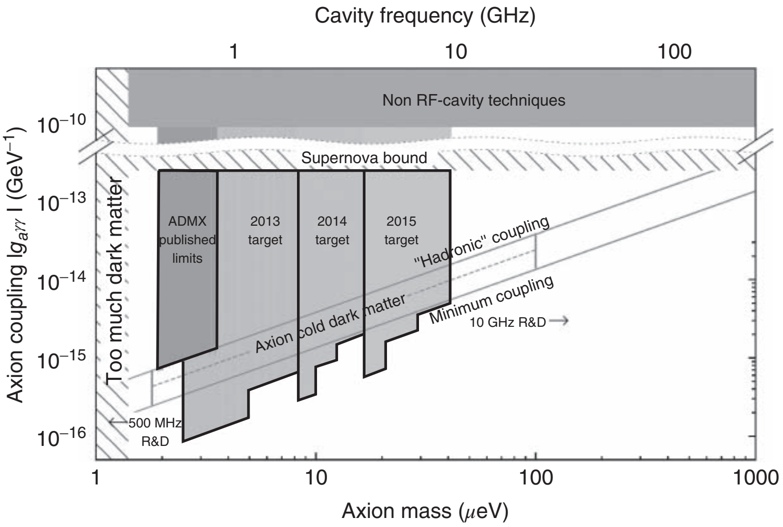


Figure 7.13 Achieved and projected sensitivity of ADMX. ADMX will soon be sensitive to even the more weakly coupled dark matter axions in the range 1–40 μeV (Reproduced with permission of [699].)

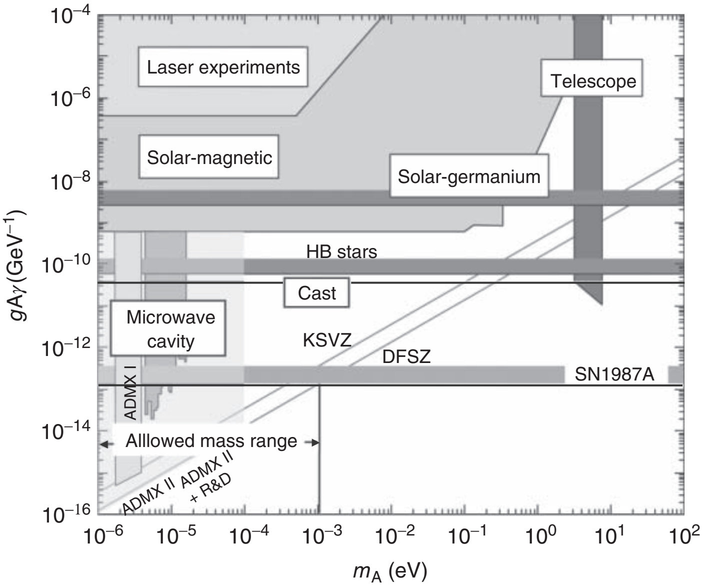


Figure 7.14 Exclusion region reported from the microwave cavity experiments RBF (Rochester-BNL-Fermilab) and UF (U. of Florida) [700–702] and ADMX [697, 698]. A local dark-matter density of 450 MeV cm^{-3} is assumed (Reproduced with permission of [703].)

Figure 7.12 illustrates a schematic of the ADMX (Axion Dark Matter eXperiment) experiment and its insert in a magnet system [696–698].

Virtual photons are provided by a 7.6-T magnetic field generated by a large superconducting solenoid with a 0.5-m diameter bore. A cylindrical copper-plated microwave cavity is embedded in the magnet bore, and the dark-matter axions passing through the cavity can resonantly convert into real microwave photons with energy $E \approx m_a^2 + (1/2)m_a^2\beta^2$. The cavity range is tunable. This experiment excluded KSVZ dark-matter axions with masses between 1.9 and 3.53 μeV (Figure 7.13). It is expected that the upgraded version of ADMX experiment will eventually cover the range 1–100 μ eV.

Summary and Prospects Figure 7.14 summarizes the present situation in the axion search. For the axion to occupy a sizable fraction of the dark matter, its mass range needs to be between 10^{-6} and 10^{-3} eV. In this range, only the microwave cavity technique can reach the KSVZ/DFSZ limit. The expected reach of the ADMX is shown in Figure 7.13.

8

Cosmology I: Big Bang Universe

8.1

Why Do We Study Cosmology?

Recent developments of particle and astroparticle physics have made boundaries of particle physics, cosmology, and astrophysics somewhat obscure. The basic notion of the Standard Model (SM) for particle physics that the unified force breaks its unification by spontaneous symmetry breakdown has offered a view to interpret the cosmic evolution in a unified way. The evolution of the expanding universe starting as a hot Big Bang can be interpreted as a sequence of phase transitions (spontaneous symmetry breaking) (see Figure 8.1). Examples of particle theory applied to cosmology include the inflation model and the Big Bang nucleosynthesis (BBN). On the other hand, the WMAP (Wilkinson microwave anisotropy probe) observation of the CMB (cosmic microwave background) in 2003 has established the SM of cosmology (referred to as the *concordance model*) and has converted it to a precision physics. It also offered a new challenge to particle physics by clarifying a new form of energy and matter, that is, the dark energy (DE) and the dark matter (DM). Together with astrophysics, they offer a vast playground for applications of the laws of microscopic physics. New particles may be found in the relics of the hot Big Bang. CMB and stars can be used as laboratories to probe fundamental interactions. We already discussed that much of the information on the neutrinos and the hypothetical axions was obtained by probing their possible role in the stellar evolutionary processes. A new frontier for particle physics is open in cosmology and astroparticle physics. In the following, we will discuss cosmology and its relation to particle physics. In this chapter, we discuss the hot Big Bang cosmology, the thermal history of the universe as a uniformly distributed fluid. In the next chapter, we discuss the structure formation (stars, galaxies, and clusters of galaxies) whose dynamics is treated as a small perturbation to the uniformly expanding universe. Without preparations on the contents of the cosmology, we cannot tackle DM and DE, which are our main interest.

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

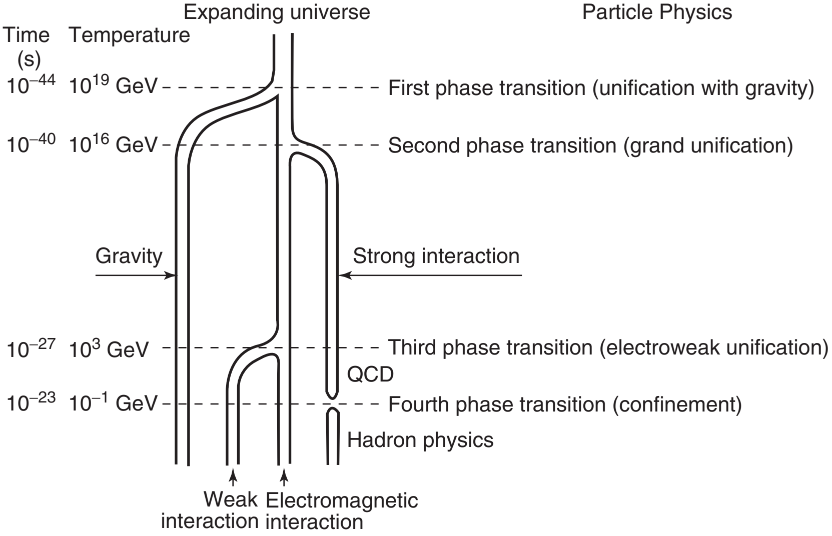

Expanding universe
Particle Physics
Time (s)
Temperature
10^{-44} 10^{19} GeV
First phase transition (unification with gravity)
10^{-40} 10^{16} GeV
Second phase transition (grand unification)
Gravity
Strong interaction
10^{-27} 10^{3} GeV
Third phase transition (electroweak unification)
QCD
10^{-23} 10^{-1} GeV
Fourth phase transition (confinement)
Hadron physics
Weak interaction
Electromagnetic interaction

Figure 8.1 Cosmic phase transition and unification of forces.

8.2
Cosmic Equation

Historically, cosmology has always been a topic of religions. Every religion has its own genesis story. At the turn of the twentieth century, it became a science when Einstein created the theory of general relativity. It is an equation of motion to solve the metric of spacetime, that is, the universe itself. The modern cosmology is constructed on the basis of three axioms:

1) Einstein's equation
2) Cosmological principle
3) Hot Big Bang hypothesis.

Einstein's equation is given by

$$\mathscr{R}_{\mu\nu} - \frac{1}{2} g_{\mu\nu} \mathscr{R} = 8\pi G_N T_{\mu\nu} \,^{1)} \tag{8.1}$$

where $c = 1$ convention is adopted as usual. The left-hand side (l.h.s.) describes dynamical evolution of the metric tensor or the curvature of spacetime. $g_{\mu\nu}$ is the metric that defines the spacetime geometry, $\mathscr{R}_{\mu\nu}$ and $\mathscr{R} = g^{\mu\nu} R_{\mu\nu}$ are the Ricci tensor and the Ricci scalar that can be derived from the metric (see Appendix E.) $T_{\mu\nu}$ is the energy–momentum tensor and G_N is the Newton's gravitational constant. The equation states that the energy distribution defines the spacetime structure. It is a differential equation of the second rank. To

---

1) More generally, a cosmological constant term $\Lambda g_{\mu\nu}$ is added to left-hand side (l.h.s.). However, it has become conventional to move it to right-hand side (r.h.s.) and interpret it as the vacuum energy density to be included as a part of $T_{\mu\nu}$.

solve the differential equation, an initial condition has to be given. The r.h.s. provides it, which is the energy–momentum tensor ($T_{\mu\nu}$) or the matter/energy distribution in the universe. The cosmological principle states that matter/energy (stars and galaxies) are distributed uniformly and isotropically in the universe.[2] The assumption is observationally true at scales over 100 Mpc.[3]

The hot Big Bang hypothesis claims that the universe started as full of hot plasma of all kinds of particles, or more precisely, the universe once has gone through a high-temperature, high-density state. The highest temperature one can think of is the Planck temperature $T_{\text{Planck}} = G_N^{-1/2} \sim 1.22 \times 10^{19}$ GeV. At this temperature, according to the general relativity, the spacetime itself is strongly curved to become of comparable size with the particle it contains, and one cannot conceive the spacetime structure as we know it. The physics is different and it is meaningless to think of a higher temperature. So let us start from the Planck temperature. All the particles and their antiparticles were in thermal equilibrium, and the number of all particle species are of the same order except for the factors that are governed by the particle's own degrees of freedom.

8.2.1
Robertson–Walker Metric

The assumption of the cosmological principle leads to the Robertson–Walker metric as the metric of the universe, which is expressed by

$$ds^2 = (dt)^2 - R^2(t)\left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right] \tag{8.2}$$

$R(t)$ has dimension of length and changes as a function of time. It represents a scale of any cosmic phenomenon but conveniently may be thought as the average intergalactic distance or size of the universe. (r, θ, ϕ) are space coordinates, which have no dimension and are referred to as the *comoving coordinates*. "k" is the curvature of the universe and, with proper normalization of R, takes value $+1, 0, -1$ depending on whether the universe is closed, flat, or open.[4] With this scaling convention, R can be considered as the radius of the whole universe in the closed universe. Today, we know that our universe is almost flat, that is, $k = 0$. But we will retain it for the sake of later discussions. We will call $d\chi = dr/\sqrt{1-kr^2}$ the comoving distance. It does not depend on time. To take cosmic expansion into account, the proper distance[5] is obtained by multiplying the scale factor ($D(t) = R(t)d\chi$). It is the physical distance measured at time t. If one adopts the normalization $R(t_0) = 1$, the comoving distance and the proper distance at present time are equal.

---

[2] If one extends the principle to include time, it becomes the stationary universe, which is denied by observation.

[3] 1 Mpc = 10^6 pc, 1 pc = 3.263 light years. The size of the entire universe is ~4000 Mpc. The best evidence is the uniformity of the CMB everywhere and in any direction.

[4] In many papers, a convention $a(t) = R(t)/R(t_0) = R(t)/R_0$ is used instead of $R(t)$. Here, $R_0 = R(t_0)$, and $t = t_0$ denotes the present time. In this case, r has to be redefined as $r \to R_0 r$, in which case it has dimension. It (i.e., the comoving distance) is the distance measured by the present scale and the curvature is modified to $k \to K = k/R_0^2$, where R_0 is the size of the present universe ~4.2 Gpc (giga parsec) = 13.7 billion light years. Later, when we discuss cosmic observables, we also adopt this convention. Then, $R_0 = 1$ and the scale factor $R(t)$ really means $R(t)/R_0$. However, we will retain the nomenclature R_0 as much as possible to avoid confusion.

At this point, we remind the reader a common misconception induced by the notion of the expanding universe. The misconception is that, after the birth of the universe, the universe expands into a vast empty space like fireworks, as is commonly depicted in the animation of expanding fireball. One must be aware that the space itself expands. Namely, the Big Bang occurred not just at one place but everywhere. All the galaxies, apart from their intrinsic local motion, do not move, and sit still in the space at the fixed coordinate, which is why the comoving coordinate system is useful. But relative distances between the galaxies are growing because of the space expansion which, to an observer, looks as if the faraway galaxies are receding with a velocity determined by the Hubble expansion rate.

An associated misconception is that a local object will expand also with time in proportion to the cosmic expansion so that one could never know if the object is expanding, because the measure by which to determine the size is also expanding. That is not true. The size of locally structured object is governed by the dynamics of the binding force. Two structured objects 1 m apart will be 2 m apart in time, but their intrinsic size does not change. Their structure including the size is determined by the electromagnetic or gravitational force. The size does not change with time even if the universe expands twofold as long as the equation of motion, that is, the physics law, does not change. Similarly, the galactic size is fixed by dynamics and does not expand. The distance between galactic clusters, on the other hand, expands because it is the size of the space between the clusters, which is not governed by the dynamics of the interactions.

Let us understand the geometry of the Robertson–Walker metric for different values of k (see Figure 8.2).

1) $k = 0$: It is obvious that, for fixed t, the Robertson–Walker metric represents the flat Euclidean space.
2) $k = +1$: If we set $r = \sin\chi$, we have $dr^2/(1 - r^2) = d\chi^2$; hence, by adopting four-dimensional polar coordinates

$$(x, y, z, u) \equiv (R\sin\chi\sin\theta\cos\phi,\ R\sin\chi\sin\theta\sin\phi,\ R\sin\chi\cos\theta,\ R\cos\chi) \quad (8.3)$$

---

[5] This is not to be confused with proper length used in special relativity, which is a length measured in the rest frame.

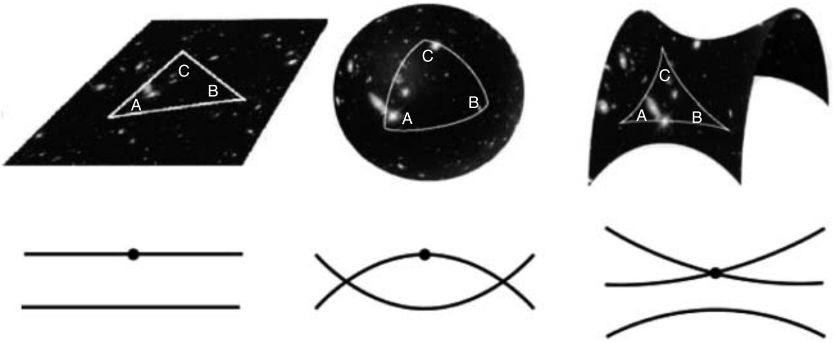


Figure 8.2 Illustration of the three kinds of space structure: flat, closed, and open universe. The three-dimensional space is depicted as a two-dimensional surface. The sum of three triangles is =, >, $< \pi$, respectively. It is also a world where one can draw 1, 0, or an infinite number of parallel lines from a point some distance apart from another line, respectively.

one obtains

$$dx^2 + dy^2 + dz^2 + du^2 = dR^2 + R^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\phi^2)] \tag{8.4}$$

$$x^2 + y^2 + z^2 + u^2 = R^2 \tag{8.5}$$

If one fixes the radius R, then $dR^2 = 0$ and Eq. (8.4) reduces to the space part of the Robertson–Walker metric. From Eq. (8.5), one sees that points $\boldsymbol{r}\,(x, y, z, u)$ represent the surface of a hypersphere with radius R embedded in the four-dimensional Euclidean space. Therefore, the three-dimensional space is closed and finite, R representing its spatial size. The comoving distance χ is represented by $\chi = \sin^{-1} r$.

To visualize the comoving distance in a closed space ($k = +1$), let us consider a sphere of unit radius, and draw a circle with center at the u-axis and radius r (Figure 8.3). Notice that we are on the surface of the sphere, and the origin of the coordinate is the North Pole. The length of the circle can be obtained by integrating over ϕ at fixed θ, which gives $2\pi r$, but the comoving distance from the North Pole to a point at $\mathbf{r} = (r, \theta, \phi)$ is a geodesic (shortest) line from the North Pole to the circle. It is given by the length of the arc which is related to r by $\chi = \sin^{-1} r$. In the vicinity of the pole, $\chi \approx r$, but in general $\chi > \sin\chi = r$.

3) $k = -1$: By setting $r = \sinh\chi$, we have $dr^2/(1 + r^2) = d\chi^2$; then

$$(x, y, z, u) \equiv (R\sinh\chi\sin\theta\cos\phi,\ R\sinh\chi\sin\theta\sin\phi,\ R\sinh\chi\cos\theta,\ R\cosh\chi) \tag{8.6}$$

one obtains

$$\begin{aligned} du^2 - dx^2 - dy^2 - dz^2 &= dR^2 + R^2[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta d\phi^2)] \\ u^2 - x^2 - y^2 - z^2 &= R^2 \end{aligned} \tag{8.7}$$

One sees that the points $\boldsymbol{r}\,(x, y, z, u)$ represent the surface of a circular hyperboloid embedded in the four-dimensional space. The comoving distance χ is the length of

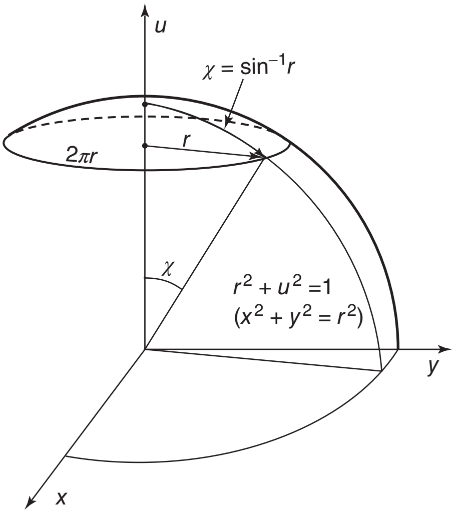


Figure 8.3 Comoving distance $\chi = \sin^{-1} r$ for $k = +1$ space.

the hyperbola from the origin to the comoving coordinate r. This time, $\chi = \sinh^{-1} r$ can be infinite, and hence the space is open.

By introducing the function

$$S_k(\chi) = \frac{1}{\sqrt{-k}} \sinh\left(\sqrt{-k}\,\chi\right) = \begin{cases} \sin\chi & k = +1 \\ \chi & k = 0 \\ \sinh\chi & k = -1 \end{cases} \tag{8.8}$$

the Robertson–Walker metric can be expressed as

$$ds^2 = dt^2 - R(t)^2\left[d\chi^2 + S_k^2(\chi)(d^2\theta + \sin^2\theta d\phi^2)\right] \tag{8.9}$$

8.2.2
Friedmann Equation

The fundamental equation for the cosmic evolution is obtained by substituting the Robertson–Walker metric into the Einstein equation, which is referred to as the *Friedmann equation.*

$$H^2(t) = \left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G_N}{3}\rho - \frac{k}{R^2} + \frac{\Lambda}{3} \tag{8.10}$$

where $\dot{R} = dR/dt$. $H(t)$ is conventionally referred to as the *Hubble constant* but is actually a function of time and represents the cosmic expansion rate. G_N is Newton's gravitational constant, ρ is the cosmic energy density including matter, radiation, and vacuum. Λ is referred to as the *cosmological constant.* It is a constant that can be positive or negative. It will be shown later that, for $\Lambda > 0$, it is equivalent to the constant vacuum energy density defined as $\rho_v = \Lambda/8\pi G_N$. When Λ is not explicitly shown in the equation, it is to be understood that it is included in the energy density. The total energy density (including that of vacuum) which makes k vanish is referred to as the *critical density*. The present ($t = t_0$) value of the critical density $\rho_c(t_0) = \rho_{c0}$ is obtained by using the present value of the Hubble constant as

$$\begin{aligned} \rho_{c0} &= \frac{3H_0^2}{8\pi G} = 1.88 \times 10^{-29} h^2 \text{gr cm}^{-3} \sim 10.5\, h^2 \text{keV cm}^{-3} \\ &= 2.78 \times 10^{11}\, h^2 M_\odot (\text{Mpc})^{-3} \end{aligned} \tag{8.11}$$

$$H_0 \equiv H(t_0) = 100\, h \,\text{km s}^{-1}\,\text{Mpc}^{-1}, \qquad h = 0.72 \pm 0.03 \tag{8.12}$$

$$M_\odot = 1.99 \times 10^{30}\,\text{kg} \qquad : \text{solar mass} \tag{8.13}$$

$$1\,\text{Mpc} = 10^6\,\text{pc},\ 1\,\text{pc (parsec)} = 3.086 \times 10^{16}\,\text{m} = 3.26 \text{ light years}. \tag{8.14}$$

The value of the critical density corresponds roughly to five protons per cubic meter or one galaxy per volume of $(1\text{ Mpc})^3$. As the size of a typical galaxy is ~30 kpc, one sees that the mass density in the galaxy is $\sim 10^4$ larger than the average density. $h = (0.72 \pm 0.03)$ represents ambiguity of present knowledge. It is conventional to leave h where the Hubble constant appears because of its long historically controversial uncertainties as well as its direct connection to observables.

Let us notice that, for $\Lambda = 0$, the Friedmann equation formally agrees with Newton's equation of motion for a particle of unit mass in the potential made by a sphere with radius R and mass M ($= 4\pi\rho R^3/3$) and total (kinetic + potential) energy $E = -k/2$.

$$\frac{1}{2}\dot{R}^2 - G_N \frac{M}{R} = E \tag{8.15}$$

This gives a convenient physical picture of the Friedmann equation. If the Hubble term is the kinetic energy, the energy density term is the potential energy and the curvature is the total energy of the universe. If the curvature term is zero, the total energy of the universe is zero, which is a very attractive idea. Equation (8.15) can serve as a convenient mnemonic to remember the Friedmann equation.

Another important equation that can be derived by applying the Robertson–Walker metric to the Einstein equation is the equation of cosmic acceleration.

$$\frac{\ddot{R}}{R} = -\frac{4\pi G_N}{3}\left(\rho + 3P\right) + \frac{\Lambda}{3} \tag{8.16}$$

where P denotes the pressure. One sees that a positive cosmological constant represents a repulsive force. Einstein's original motivation for introducing the cosmological constant was to set the net force to vanish to realize a stationary universe because he did not know that the universe was expanding. The first term in the parenthesis of Eq. (8.16) represents Newton's law of gravity. This can be shown by rewriting the equation in terms of $M = \rho(4\pi R^3/3)$. The second pressure term is added by the general relativity. The pressure can be negative and, if an equality $w \equiv P/\rho < -1/3$ holds, the net force will be repulsive for $\Lambda = 0$. In fact, using Eqs. (8.10) and (8.16), and the fact that a constant vacuum density means

$P = -\rho$,⁽⁶⁾ it is easy to show that the cosmological constant is equivalent to the constant vacuum energy defined by

$$\rho_{\text{vac}} = \frac{\Lambda}{8\pi G_N} \tag{8.18}$$

In the following, the cosmological constant term is to be understood as a part of the energy density unless explicitly stated.⁽⁷⁾ By combining Eqs. (8.10) and (8.16), one obtains another relation

$$\dot{\rho} + 3H(\rho + P) = 0 \tag{8.19}$$

Equation (8.19) can be re-expressed in the form

$$\frac{d}{dt}(\rho R^3) + P\frac{d}{dt}R^3 = 0 \tag{8.20}$$

which can be compared with the first law of thermodynamics with $dQdU + PdV = 0$; namely the cosmic entropy is conserved. This is a consequence of cosmological principle which forbids the net heat flow in any sampled volume in the comoving coordinate frame. Note also that the uniform isotropic universe can be considered as a perfect fluid, for which the energy–momentum tensor takes the form⁽⁸⁾

$$T_{\mu\nu} = -Pg_{\mu\nu} + (\rho + P)u_\mu u_\nu \tag{8.21}$$

where $u_\mu = dx_\mu/d\tau = (1/\sqrt{1-v^2})dx_\mu/dt$ is the four velocity fields in units of the light velocity c.

When particles interact and change their identity, the first law of thermodynamics is extended to

$$dU = TdS - PdV + \sum_i \mu_i dN_i \tag{8.22}$$

where $U = \rho V$ is the internal energy, S the total entropy in a volume V and μ_i is the chemical potential of particle species i. Equation (8.20) tells that (8.22) is equivalent to

$$\frac{dS}{dt} = -\frac{1}{T}\sum_i \mu_i \frac{dN_i}{dt} \tag{8.23}$$

If the chemical potential can be neglected, or if there is no particle production/annihilation, the entropy is conserved. If the particle numbers are changing

---

6) The first law of thermodynamics tells us $dU = d(\rho V) = dQ - PdV$. For an isolated volume, $dQ = 0$. Therefore, the constant energy density means

$$\frac{d\rho}{dV} = -\frac{\rho + P}{V} = 0 \quad \rightarrow \quad P = -\rho \tag{8.17}$$

7) The true identity of the vacuum energy, generically referred to as the *dark energy*, is a hot theme in cosmology. It could be a dynamical object, in contrast to the static cosmological constant. It will be discussed in more detail in Chapter 11. As of now, there is no observational evidence that it is something other than the cosmological constant.

8) Real fluids are sticky. The perfect fluids have no shear stresses, viscosity or heat conduction.

but are in chemical equilibrium in the reaction $1+2 \leftrightarrow 3+4$, the chemical potentials are constrained by

$$\mu_1 + \mu_2 = \mu_3 + \mu_4 \tag{8.24}$$

Each time the reaction $1+2 \leftrightarrow 3+4$ occurs, the particles 1 and 2 disappear and 3 and 4 appear. Consequently, the number changes are related by $dN_1 = dN_2 = -dN_3 = -dN_4$, which results in $\sum_i \mu_i dN_i = 0$. If other reactions occur, the same conclusion holds. In other words, the entropy is conserved in reactions in chemical equilibrium. The entropy is produced only in a special case when particle numbers change in thermal nonequilibrium conditions.

The cosmic equations contain three variables, R, ρ, and P. As only two of the equations Eq. (8.10), (8.16), and (8.19) are independent, we need another equation to solve them as a function of time. It is provided by the equation of state to relate the energy density to the pressure.

$$P = w\rho \qquad w = \begin{cases} 1/3 & \text{radiation}^{9)} \\ 0 & \text{matter (nonrelativistic (NR) particle)} \\ -1 & \text{vacuum energy.} \end{cases} \tag{8.25}$$

Generally, the equation of the state variable w is a function of time or, equivalently, the redshift "z" (to be defined soon). However, we treat it as a constant except when we discuss the dark energy in Chapter 11. As the equations are differential equations of the second order (or two differential equations of the first order), two boundary conditions are necessary which are supplied by ρ_0 and the curvature k.

8.3
Expanding Universe

8.3.1
Redshift of Light

As the light path is defined by $ds^2 = 0$ or, equivalently, $d\eta^2 \equiv dt^2/R^2(t) = d\chi^2$, which means

$$\eta(t) - \eta(0) = \int_0^t \frac{dt}{R(t)} = \int_0^r \frac{dr}{\sqrt{1-kr^2}} \tag{8.26}$$

Where we set $\theta = \phi = 0$ without loss of generality. Suppose one receives a light at $r = 0$, at time $t = t_0 \sim t_0 + dt_0$, which was emitted at $r = r_1$, $t = t_1 \sim t_1 + \delta t_1$ in the comoving coordinate system, the r.h.s. of Eq. (8.26) does not depend on time and

[9] See Eq. (8.43). Generally, the equation of state is a function of two independent thermodynamic variables, but with the entropy conservation constraint it can be expressed in this form. For nonrelativistic particles, the equation of state is $P = nkT : kT \ll m$ and $P \simeq 0$ is a good approximation.

we obtain $\delta\eta(t_0) = \delta\eta(t_1)$. Consequently,

$$\frac{\delta t_0}{R(t_0)} = \frac{\delta t_1}{R(t_1)} \tag{8.27}$$

Writing the wavelength and frequency of the light as λ and ν, and using $1/\nu$ as a measure of δt, we have

$$\begin{aligned} &\nu_1 R(t_1) = \nu_0 R(t_0) \equiv \nu_0 R_0 \\ \therefore \quad &\frac{\nu_1}{\nu_0} = \frac{\lambda_0}{\lambda_1} \equiv 1 + z = \frac{R_0}{R(t_1)} > 1, \quad z \equiv \frac{\lambda_0 - \lambda_1}{\lambda_1}. \end{aligned} \tag{8.28}$$

It means that an observed light from faraway galaxies has longer wavelength than that at its emission. That is, the light is always shifted toward red and the above equation defines the amount of redshift z.[10]

8.3.2
Redshift of Particles

We can also show that particles are redshifted, as the universe expands; that is, its observed momentum is smaller than that it had at the time of emission. Suppose the particle with energy–momentum (E, p) flies δx in the comoving coordinate system during time δt; then the actual flight length is given by

$$R(t)\delta x = v\delta t, \qquad v = \frac{p}{E} \tag{8.29}$$

In the expanding universe, an observer has relative velocity $\beta = d(R\delta x)/dt = \dot{R}\delta x$ to the particle source. The momentum seen by the observer is Lorentz-transformed to

$$p' = p - \beta E = p\left(1 - \frac{\beta}{v}\right) \tag{8.30}$$

Therefore, the momentum shift is given by

$$\frac{\delta p}{p} = -\frac{\beta}{v} = -\frac{\dot{R}\delta x}{v} = -\frac{\dot{R}}{R}\delta t \tag{8.31}$$

which leads to

$$pR = \text{constant} \quad \rightarrow \quad \frac{p}{p_0} = \frac{R_0}{R(t)} = 1 + z \tag{8.32}$$

8.3.3
Cosmic Parameters

Cosmic parameters are inputs to the Friedmann equations, and their present values can be determined by observations. It is conventional to express the energy density in units of the critical density ρ_c, namely

---

[10] When the galaxy is moving in the comoving coordinate system (i.e., relative to the Hubble expansion), its velocity (referred to as the *peculiar velocity*) perturbs the value of the redshift. It is generally small and is neglected in most of our discussions. Our neighbor, the Andromeda, is blue-shifted because of local gravitational force.

$$\begin{aligned}\Omega_i(t) &\equiv \frac{\rho_i(t)}{\rho_c(t)}, \qquad i = m\,(\text{matter}),\ r\,(\text{radiation}),\ \Lambda(\text{vacuum})\\ \rho_c(t) &= \frac{3c^2H^2(t)}{8\pi G_N}.\end{aligned} \tag{8.33}$$

Let us rewrite the Friedmann equation in terms of the cosmic parameters. Dividing both sides of Eq. (8.10) by ρ_c and defining $\Omega_k \equiv -kc^2/(R^2H^2(t))$, the Friedmann equation takes a form which is very transparent intuitively.

$$\Omega(t) + \Omega_\Lambda = \Omega_m(t) + \Omega_r(t) + \Omega_\Lambda = 1 - \Omega_k(t) \tag{8.34}$$

The equation of cosmic acceleration Eq. (8.16) becomes

$$q(t) \equiv -\frac{\ddot{R}/\dot{R}}{\dot{R}/R} = \Big(1 + 3w(t)\Big)\frac{\Omega(t)}{2} - \Omega_\Lambda \tag{8.35}$$

where $q(t)$ is referred to as the *deceleration parameter*. The vacuum energy density as opposed to the cosmological constant need not be a constant, as it could appear as a guise of potential energy of a certain field (see Section 11.2, Quintessence model). But, in the following we will assume Ω_Λ =constant unless otherwise stated. Equation (8.34) says that it is the sum of energy densities that determines the overall sign of the curvature. However, there is a strong theoretical reason (the inflation model: see Section 8.8) that the curvature of the universe vanishes. Then Eq. (8.34) says that the total energy density equals the critical density.

If we accept the inflation model, Eq. (8.34) at present time ($t = t_0$) becomes

$$\Omega_{m0} + \Omega_{r0} + \Omega_\Lambda = 1 \tag{8.36}$$

Ω_Λ is really $\Omega_{\Lambda 0} = \Omega_\Lambda(t = t_0)$, but suffix 0 is omitted because of our assumption Ω_Λ = constant. At present ($t = t_0$), the pressure P can be neglected. Therefore

$$q_0 = \frac{\Omega_{m0}}{2} - \Omega_\Lambda \tag{8.37}$$

Ω_{r0} is determined from the CMB accurately and turns out to be negligible (see Eq. (8.52)). Ω_{m0} can be measured by counting the abundance of galaxies and dark matter, or estimated from light element abundance using the nucleosynthesis argument discussed in Section 8.6.5. The deceleration parameter q_0, or equivalently $\Omega_{\Lambda 0}$, can be determined from the observed cosmic acceleration rate (see Section 11.1.1). The curvature parameter Ω_{k0} can be determined by comparing the predicted size of the CMB horizon to its view angle (see Section 9.2.6). Therefore, Eq. (8.34) at present time can be determined with one constraint, and the validity of the Friedmann equation (and by extension the general relativity) can be tested. The cosmic parameters and their present values are listed in Table 8.1. [7,704]

Table 8.1 Cosmic parameters.[a]

		Measured value at $t = t_0$
Hubble parameter	$H = \dot{R}/R$	$H_0 = 100\,h\ \text{km s}^{-1}\ \text{Mpc}^{-1}$
		$h = 0.72 \pm 0.03$
Critical density	$\rho_c = \dfrac{3c^2H^2}{8\pi G_N}$	$\rho_{c\,0} = 1.88 \times 10^{-29} h^2 \text{gr cm}^{-3}$
Density parameter	$\Omega = \rho/\rho_c = \Omega_{\text{rad}} + \Omega_m$	
Radiation density		$\Omega_{\text{rad}\,0} = 7.73 \pm 0.4 \times 10^{-5}$
Matter density		$\Omega_{m\,0} = 0.26 \pm 0.02$
(Baryon density[b])		$\Omega_{b\,0} = 0.045 \pm 0.03$
Dark energy parameter	$\Omega_\Lambda = \dfrac{\Lambda c^2}{3H_0^2} = \dfrac{\rho_v}{\rho_c}$	0.74 ± 0.03
Curvature parameter	$\Omega_k = -\dfrac{kc^2}{R^2H_0^2}$	0.006 ± 0.006
Deceleration parameter	$q = -\dfrac{\ddot{R}/\dot{R}}{\dot{R}/R} = -\dfrac{\ddot{R}}{RH^2}$	$q_0 = -0.61 \pm 0.03$

[a] Note: Recent Planck data gives $h = 0.6780 \pm 0.0077$, $\Omega_b h^2 = 0.02214 \pm 0.00024$ $\Omega_{m0} h^2 = 0.1187 \pm 0.0017$, $\Omega_\Lambda = 0.692 \pm 0.010$ [674].

[b] Part of the matter density.

Dark matter and ΛCDM model: The table shows that the dark energy (i.e., the vacuum energy) occupies 74% of the total cosmic energy and the matter rest of 26%. Notice, however, that the baryon density is a fraction of the matter density and occupies only 4.5% of the total. In other words, contribution of the SM particles is only a tiny fraction of the total energy budget. The rest of the matter is an unknown object referred to as *dark matter (DM)*. It is known that it is stable and does not interact strongly or electromagnetically. Its existence can be detected only through gravitation or other indirect methods. A detailed discussion of DM and DE will be presented in Chapters 10 and 11. However, their presence in the background cosmological environment is quite apparent throughout in discussing the contents of the expanding universe. The cold dark matter (CDM), which is nonrelativistic at the time of matter-radiation equality (see Section 8.7.1) plays a particularly important role in formation of the large-scale structure, which is the subject of the next chapter. The ΛCDM model, which takes into account the cosmic parameters in Table 8.1 and some others, is now considered as the standard model of cosmology. The nomenclature Λ distinguishes it from old models that ignored the vacuum energy contribution. The concordance model, as it is also called, was established around 2003 after analysis of the WMAP precision data [705].

8.4
Thermal Universe

8.4.1
Thermodynamics

The Friedmann equation Eq. (8.10) and the energy conservation law (8.20) are the two fundamental equations in cosmology. By adding the equation of state to relate the energy density ρ to the pressure P, essential features of cosmic evolution can be deduced. The number density of a particle with energy E in thermal equilibrium is given by the Planck formula

$$dn = \frac{g}{(2\pi)^3} \frac{d^3p}{\exp\left[\frac{E-\mu}{T}\right] \pm 1}. \tag{8.38}$$

where g is the particle's number of degrees of freedom (= 2 for the photon and ordinary fermion, the antifermion is counted separately) and μ is the chemical potential. The temperature T can be converted into energy by multiplying the Boltzmann constant $T \to k_B T$. In the following, we always express the temperature in energy, unless otherwise noted, and drop k_B.

When the particle species i is in thermal equilibrium, its number density n_i and energy density ρ_i in the relativistic ($T \gg m_i$) and nondegenerate (ND) ($T \gg |\mu|$) conditions are given by

$$n_i = \int dn_i = g_i \frac{\zeta(3)}{\pi^2} T^3 \times \begin{cases} 1 & : \text{Bose–Einstein (BE) statistics} \\ 3/4 & : \text{Fermi–Dirac (FD) statistics} \end{cases} \tag{8.39a}$$

$$\rho_i = \int E_i dn_i = g_i \frac{\pi^2}{30} T^4 \times \begin{cases} 1 & : \text{(BE)} \\ 7/8 & : \text{(FD)} \end{cases} \tag{8.39b}$$

$\zeta(3) = 1.20205$ is the zeta function of degree 3. In many cases, NR ($T \ll m_i$) and ND approximations will be used.

$$\left.\begin{aligned} n_i &= g_i \left[\frac{mT}{2\pi}\right]^{\frac{3}{2}} e^{\frac{\mu-M}{T}} \\ \rho_i &= n_i m + \frac{3}{2} n_i T \end{aligned}\right\} \quad \text{NR, ND} \tag{8.40}$$

The photon's chemical potential can be assumed to vanish. A justification comes from the observed cosmic microwave spectrum, which agrees well with the Planck formula with $\mu = 0$ to a high precision. Another argument comes from the fact that the net chemical potential in the early universe can be set to zero, because the asymmetry between particles and antiparticles ($\gamma + \gamma \leftrightarrow e^+ + e^-$) is very small (see discussions in Section 8.6.5).

The present temperature of the CMB radiation is measured to be $T_{\gamma,0} = 2.725 \pm 0.002\,K$ [7]. The present number of photons is essentially that of CMB, and by using Eq. (8.39a) it is calculated to be

$$n_{\gamma 0} = 2 \times \frac{\zeta(3)}{\pi^2} T^3 \bigg|_{T=T_0} = 410.5 \text{ cm}^{-3} \times \left(\frac{T_0}{2.725\,K}\right)^3 \tag{8.41}$$

The pressure P is given by the perpendicular component to a plane of the total momentum per unit time per unit area.

$$P = \int p\cos\theta v\cos\theta\, dn = \frac{1}{3}\int \frac{dE}{dp} p\, dn \tag{8.42}$$

which gives the equation of state.

$$P_i = \frac{1}{3}\rho_i \qquad \text{radiation or extremely relativistic (ER) particles} \tag{8.43a}$$

$$= n_i T \qquad \text{non-relativistic (NR) nondegenerate (ND) particles.} \tag{8.43b}$$

Entropy S and the first law of thermodynamics are related by

$$dU = TdS - PdV \tag{8.44a}$$

$$\begin{aligned} TdS &= dU + PdV = d[\rho(T)V] + P(T)dV \\ &= V\frac{d\rho(T)}{dT}dT + [\rho(T) + P(T)]dV, \end{aligned} \tag{8.44b}$$

$$\therefore \quad \frac{\partial S}{\partial T} = \frac{V}{T}\frac{d\rho(T)}{dT}, \quad \frac{\partial S}{\partial V} = \frac{\rho(T) + P(T)}{T}. \tag{8.44c}$$

From the integrability condition $\partial^2 S/\partial V\partial T = \partial^2 S/\partial T\partial V$, we obtain

$$\frac{dP(T)}{dT} = \frac{\rho + P}{T} \tag{8.45}$$

Using Eqs. (8.45) and (8.44), one can derive

$$dS = d\left(\frac{\rho + P}{T}V\right) \tag{8.46}$$

In thermal equilibrium, $dS = 0$, and therefore the entropy S or $[(\rho + P)/T]V$ is conserved. The entropy and energy densities of relativistic particles in thermal equilibrium are given by

$$s = \frac{\rho + P}{T} = \frac{4}{3}\frac{\rho_\gamma}{T} \overset{(8.39b)}{=\!=} \frac{4}{3}\frac{\pi^2}{30}T^3 \times \begin{cases} 1 & \text{boson} \\ \frac{7}{8} & \text{fermion.} \end{cases} \tag{8.47a}$$

For mixtures of many particle species, the entropy and energy density can be expressed in terms of the photon temperature.

$$s = g_s(T)\frac{2\pi^2}{45}T^3, \tag{8.47b}$$

$$\rho_{\text{rad}} = g_*(T)\frac{\pi^2}{30}T^4, \tag{8.47c}$$

$$g_s(T) = \sum_{\text{boson}} g_B\left(\frac{T_B}{T}\right)^3 + \frac{7}{8}\sum_{\text{fermion}} g_F\left(\frac{T_F}{T}\right)^3, \tag{8.47d}$$

$$g_*(T) = \sum_{\text{boson}} g_B\left(\frac{T_B}{T}\right)^4 + \frac{7}{8}\sum_{\text{fermion}} g_F\left(\frac{T_F}{T}\right)^4 \tag{8.47e}$$

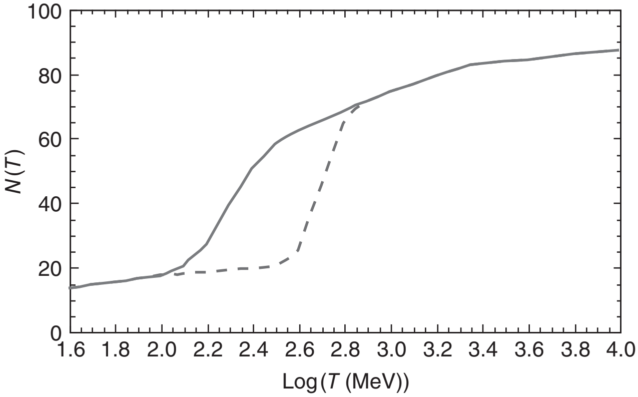


Figure 8.4 Effective particles degree of freedom in the Standard Model [$SU(3) \times SU(2) \times U(1)$]. The real and dashed lines assume $\lambda_{\mathrm{QCD}} = 150$ and 450 MeV, respectively.

where T_B, T_F are temperatures of the bosons and fermions, respectively. However, for $T \gg m_e$, we may set $g_s(T) = g_*(T) = N(T)$ where

$$N(T) = \sum_{\text{boson}} g_B + \sum_{\text{fermion}} \frac{7}{8} g_F.^{11)} \tag{8.48}$$

$N(T)$ is the effective number of degrees of freedom at temperature T. It is given by summing all the relativistic particles in thermal equilibrium and is a function of the temperature. For instance at $m_e \ll T \ll m_\mu$, the relativistic particles are photons, electrons, and neutrinos in the SM[12] and $N(T) = 2\,(\gamma) + 7/8 \times [2\,(e^-) + 2\,(e^+) + 3 \times 2(\nu_i\; i = 1-3)] = 43/4 = 10.75$. Figure 8.4 shows the temperature dependence of $N(T)$ in the SM. The total entropy in a comoving volume is kept constant, which leads to

$$S \sim sR^3 \sim g_s(T)(TR)^3 \sim \text{constant} \quad \rightarrow \quad TR \propto g_s(T)^{\frac{1}{3}} \tag{8.49}$$

As $g_s(T) \simeq$ constant except for the very early universe, we have the important relation

$$TR = \text{constant}. \tag{8.50}$$

As can be seen from Eqs. (8.39a) and (8.47), the entropy density is proportional to the number of relativistic particles. Consequently, the present entropy density is practically determined by the number density of the photon and the neutrino. For $T \ll m_e$ (i.e., now), the neutrino temperature is different from the photon

---

11) We will use $N(T)$, $g_s(T)$, $g_*(T)$ interchangeably where there is no fear of confusion. However, the distinction is important in matter dominated era in which case we explicitly write down the formula.

12) In the non-SM, any particles with mass $\ll m_\mu$ have to be counted also.

temperature ($T_\nu = (4/11)^{1/3} T_\gamma$, see Eq. (8.112)).[13]

$$s = \frac{4}{3}\frac{\pi^2}{30}\left(2 + \frac{21}{4}\left(\frac{T_\nu}{T_\gamma}\right)^3\right)T^3 = \frac{\pi^4}{45\zeta(3)}\left(2+\frac{21}{11}\right)n_\gamma = 7.0395 n_\gamma \tag{8.51}$$

Similarly, today's total radiation energy is given by

$$\begin{aligned}\rho_{\text{rad}} &= g_*(T_\gamma)\frac{\pi^2}{30}T_\gamma^4 = \left[2 + 3\times 2\times\frac{7}{8}\left(\frac{T_\nu}{T_\gamma}\right)^4\right]\frac{\pi^2}{30}T_\gamma^4 \simeq 1.6813\rho_\gamma \\ \Omega_{r0} &= \frac{\rho_{\text{rad}}}{\rho_c} = 4.154\times 10^{-5}\left(\frac{T_0}{2.725}\right)^4 h^{-2} \approx 7.734\times 10^{-5}.\end{aligned} \tag{8.52}$$

The amount of light from the stars is negligible. Equation (8.52) means that the radiation energy at the present time can be neglected. However, in the early universe ($T/T_0 = R_0/R = 1+z \gtrsim 3000$), it plays an essential role.

8.4.2
Radiation and Matter Dominance

In the early universe ($1+z \gtrsim 3000$), contributions to the energy density of the universe are mainly from relativistic particles. It is referred to as the *radiation-dominated universe*. When the main contribution comes from nonrelativistic particles, it is referred to as the *matter-dominated universe*. In the far future, the cosmological constant term, if it is truly a constant, eventually dominates, which we will refer to as the Λ-*dominated universe*.

Combining Eqs. (8.19) and (8.25) and for constant w, we can express the energy density as a function of the scale.

$$\rho R^{3(1+w)} = \text{constant} \tag{8.53}$$

In the radiation-dominated universe, $P = \rho/3$ or $w = 1/3$, and in the matter-dominated universe, $P \sim 0$ or $w = 0$. Then from Eqs. (8.39b) and (8.50), one can immediately derive

$$\rho(t) \sim \begin{cases} \frac{1}{R^4(t)} \sim T^4 & : \text{ radiation dominated} \\ \frac{1}{R^3(t)} \sim T^3 & : \text{ matter dominated} \\ = \text{constant} & : \Lambda \text{ dominated}^{14)} \end{cases} \tag{8.54}$$

Then the curvature term ($\sim R^{-2} \sim T^2$) or cosmological constant term (~constant) in the Friedmann equation Eq. (8.10) can be neglected in the early epoch. For an era when only one component dominates, the Friedmann equation Eq. (8.10) can

---

13) If the neutrinos are massive, they may no longer be relativistic. But the same reasoning by which one can define the neutrino temperature (see boxed paragraph after Eq. (8.112)) constrains its spectrum to be given by that of the relativistic particles.

14) The universe that has only the cosmological constant term without matter or radiation content is referred to as *de Sitter universe*. If $\Lambda < 0$, it is referred to as the *anti-de Sitter universe*.

easily be solved to give

$$R(t) \sim T^{-1} \sim \begin{cases} \sim t^{\frac{1}{2}} & : \text{radiation dominated} \\ \sim t^{\frac{2}{3}} & : \text{matter dominated} \\ \sim e^{\sqrt{\frac{\Lambda}{3}}t} \equiv e^{H_\Lambda t} & : \Lambda \text{ dominated.} \end{cases} \tag{8.55}$$

This is a very important relation which will be used frequently in the following section. To emphasize, the scale of the universe expands $R \sim 1/T$ in inverse proportion to the temperature, but the expansion rate is different at each epoch; $\sim t^{1/2}$ in the radiation-dominated era; $\sim t^{2/3}$ in the matter dominated era; and exponential ($\sim e^{H_\Lambda t}$) in the Λ-dominated era. The Hubble constant will also have a characteristic time dependence at each epoch.

$$H(t) = \frac{\dot{R}}{R} = \begin{cases} (2t)^{-1} & : \text{radiation dominated} \\ (3t/2)^{-1} & : \text{matter dominated} \\ H_\Lambda = \sqrt{\frac{\Lambda}{3}} & : \Lambda \text{ dominated} \end{cases} \tag{8.56}$$

H^{-1} is referred to as the *Hubble time*.

8.4.3
Time versus Temperature

The epoch that concerns particle physicists most is the radiation-dominated era. We calculate a few formulas that will be useful for later discussions. Using Eq. (8.39b) and the Friedmann equation, the cosmic energy density ρ_{rad} is given by

$$\rho_{\text{rad}} = g_*(T)\frac{\pi^2}{30}T^4 = \frac{3}{32\pi G_N}t^{-2} \tag{8.57}$$

where g_* is given by Eq. (8.47e). The last equality can be derived by substituting Eq. (8.56) into the Friedmann equation [Eq. (8.10)] and setting $k = \Lambda = 0$. In an epoch where $N(T)$ does not change appreciably (e.g., $N(T) \simeq 10.75$ for $m_e < T < m_\mu$, $N(T) = 3.36$ for $T \ll m_e$),

$$t = \left[\frac{90}{32\pi^3 g_*(T) G_N}\right]^{\frac{1}{2}} \frac{1}{T^2} \simeq 0.30\frac{M_{\text{Pl}}}{T^2\sqrt{g_*(T)}}, \tag{8.58}$$

$$\begin{aligned} T &= \left[\frac{90}{32\pi^3 g_*(T) G_N}\right]^{\frac{1}{4}} t^{-\frac{1}{2}} \\ &= \frac{1.55\,\text{MeV}}{g_*^{\frac{1}{4}}(T)\sqrt{t}} \simeq \frac{0.95\,\text{MeV}}{(t/\text{s})^{\frac{1}{2}}} \simeq \frac{1.09\times 10^{10}\,K}{(t/\text{s})^{\frac{1}{2}}} \qquad \text{for } T \lesssim m_e \end{aligned} \tag{8.59}$$

where $M_{\text{Pl}} = 1/\sqrt{G_N} \simeq 1.22 \times 10^{19}$ GeV is the Planck mass. Another useful relation at the same epoch is

$$\begin{aligned} H(t) &= \alpha T^2 \\ \alpha &= \sqrt{\frac{4\pi^3}{45} g_*(T) G_N} \simeq \frac{1.66\sqrt{g_*(T)}}{M_{\text{Pl}}}. \end{aligned} \tag{8.60}$$

8.4.4
Overview of Thermal History

The dynamical evolution of the universe is dictated by the Friedmann equation expanding with the rate given by $H(t)$.

As the universe expands, the cosmic temperature goes down and all particles are redshifted. As the temperature goes down, various kind of phase transitions occur. Particles once in thermal equilibrium are decoupled from others at a certain time determined by the particle's interaction rate Γ and the Hubble expansion rate H. The particle's reaction rate Γ is given by

$$\Gamma \approx n v \sigma \tag{8.61}$$

where n is its number density, v its average velocity determined by the temperature, and σ its reaction cross section. Denoting the decoupling temperature as T_{dc}, it is determined by the equation

$$H(T_{\text{dc}}) \approx \Gamma^{\;15)} \tag{8.62}$$

when the cosmic expansion rate equals and exceeds the reaction rate. This can be understood as follows: $1/\Gamma$ is the mean free path of a particle. If it is longer than the particle horizon $1/H$ (see Section 8.5.2), interactions do not happen, and hence the particle is decoupled.

Before discussing each incident, we will have a bird's eye view of the thermal history to help understand the arguments describing various incidents in the cosmic history.

$T \gg 1$ MeV After the birth of the universe, it is believed to have gone through a period of rapid expansion (exponential growth) referred to as the *inflation*. It has wiped out any traces of past cosmic history. The universe is cooled down by the inflation, and is reheated to make the universe filled with hot plasma again, which is the beginning of the Big Bang.

The inflation is believed to have happened during the phase transition at the GUT's (grand unified theory's) temperature ($T \sim 10^{16}$ GeV) when the strong

---

15) For an accurate estimate of the decoupling time, it is necessary to make a statistical treatment of nonequilibrium processes using Boltzmann equations. It is known that the results obtained using this simple criterion (known as *Gamow's criterion*) agree well with detailed calculations.

interaction separates from the electroweak interaction (see Figure 8.1). At $T \sim G_F^{-1/2} \sim 250\,\text{GeV}$, the electroweak phase transition occurred, and at $T \sim \Lambda_{\text{QCD}} \sim 200$ MeV the quark confinement phase transition occurred. It is believed that a tiny excess (1 in 10^{10}) of the quark and/or lepton was created at the GUTs era which may have survived or been converted to the baryon excess at the electroweak phase transition (baryogenesis). The baryons that have survived the phase transitions constitute bulk of the matter universe we observe today.

After the confinement phase ($T \sim 100 - 200$ MeV), the quarks fuse themselves to create nucleons which, in turn, are synthesized to form nuclear compounds. But they are immediately destroyed by the overwhelming numbers of photons and neutrinos.

$T = 1 \sim 0.1$ MeV At time $t \sim 1$ s ($T \sim 1$ MeV), most of the particle species have already disappeared by pair annihilation, and the universe contains only photons, neutrinos, and a small number of nucleons and a matching number of electrons. They were in thermal equilibrium. As the universe expands, the neutrino decouples at this time, and throughout the rest of the cosmic history its only role is to contribute to the total energy budget of the universe. At $t = 3$ min ($T \sim 0.1$ MeV), photodisintegration of the deuterium ceases and light elements including hydrogen, helium, and lithium are synthesized (referred to as the *Big Bang nucleosynthesis*, BBN), which constitute bulk of the ordinary matter as they exist today. At this stage, they are in the plasma state. They cannot form neutral atoms (hydrogens and heliums) because they are immediately destroyed by the photons.

Recombination At $t \simeq 380\,000$ years ($T \sim 0.3$ eV), the photon can no longer sustain thermal equilibrium with protons (and helium nuclei) and electrons. Neutral hydrogens and heliums are formed. The time is referred to as the *recombination*. After completion of the recombination, there are no charged particles to scatter the photons. The photon travels freely (photon decoupling), that is, the universe becomes transparent to the photons. The part of the cosmic sphere at the time of decoupling is referred to as the *surface of last scattering*. These photons are observed today as the cosmic electromagnetic wave noise covering the whole sky, which is referred to as *CMB (cosmic microwave background radiation)*. Their spectrum exactly reproduces the Planck's black body radiation formula at the temperature $T_\gamma = 2.725$ K, which has the microwave frequency of 160.2 GHz, corresponding to 1.9 mm wavelength.

Matter-Radiation Equality An event which is not in the realm of thermodynamics but which later will have an important effect on the structure formation of the universe has happened at $t \sim 75\,000$ years ($T \sim 0.76$ eV), shortly before the decoupling. It is the time when the matter density exceeded that of radiation and local gravitational fall commenced. One refers to this as the transition from radiation dominance (RD) to matter dominance (MD). This is the beginning of the structure formation, which has developed into stars, galaxies, and clusters of galaxies. At about 5.9 Gyrs (Giga years), the dark energy became dominant component of the cosmic energy and acceleration of the cosmic expansion began again.[16] The future cosmic destiny critically depends on the property of the dark energy, which has turned the decelerating expansion to an accelerating expansion. It will be discussed in detail in Chapter 11, but its effects are minimal for the phenomena discussed in this chapter because the dark-energy-dominated era has begun very recently. For this reason, most of the cosmic expansion during the span of 13.7 billion years can be referred to as *matter dominated*.

We list some observational evidences of the standard Big Bang cosmology.

1) The distribution of the CMB is uniform and isotropic to a precision of 10^{-5}.
2) The universe is structured (distribution of galaxies are clumpy) at various levels, but at larger scale than ~100 Mpc, distributions of galaxies and sources of radio waves are uniform.
3) The farther the galaxy, the larger its redshift.
4) The standard Big Bang cosmology reproduces cosmic abundance of light elements (H/D/He/Li) and the large-scale structure of galaxies quite well.

The cosmological principles (1) and (2) and the expanding universe (3) are the foundations of the Friedmann equation and the existence of CMB (1) and (4) support the notion of the Big Bang cosmology.

8.5
Cosmic Distance, Horizon

8.5.1
Distance

Comoving and Proper Distance Sometimes it is convenient to express the Robertson–Walker metric in terms of the comoving distance χ and conformal time defined by $d\eta = dt/R(t)$.

$$ds^2 = R^2(\eta)\left[d\eta^2 - \left\{d\chi^2 + S_k^2(\chi)(d\theta^2 + \sin^2\theta d\phi^2)\right\}\right] \tag{8.64}$$

In the following discussions, we can set $\theta = \phi = 0$ without loss of generality. The relation between the comoving radial coordinate r and comoving distance is given by

---

[16] The transition time t_A from deceleration to acceleration can be determined from Eq. (8.37).

$$q = (1+z_A)^3\frac{\Omega_{m0}}{2} - \Omega_\Lambda = 0, \quad (1+z_A)^3 = \left(\frac{R_0}{R_A}\right)^3 \overset{(8.55)}{=\!=} \left(\frac{t_0}{t_A}\right)^2 = \frac{2\Omega_\Lambda}{\Omega_{m0}} \tag{8.63}$$

which gives $z_A = 0.76$, $t_A = 5.9$ Gyrs.

$$r = S_k(\chi) = \frac{1}{\sqrt{-k}}\sinh(\sqrt{-k}\,\chi) = \begin{cases} \sin\chi & k = +1 \\ \chi & k = \ \ 0 \\ \sinh\chi & k = -1 \end{cases} \tag{8.65a}$$

$$\chi = \int_0^r \frac{dr}{\sqrt{1-kr^2}} = \begin{cases} \sin^{-1} r & k = +1 \\ r & k = \ \ 0 \\ \sinh^{-1} r & k = -1 \end{cases} \tag{8.65b}$$

In cosmological equations, r appears frequently, but to connect it to the observables, one uses the light path relation $ds^2 = (cdt)^2 - R(t)^2 d\chi^2 = 0$. That is, χ is the distance the light travels from its emission to an observer at r in the comoving coordinate system. The physical (proper) distance is obtained by multiplying χ by the scale factor $R(t)$.

For the light, the proper comoving distance $d\chi$ agrees with the conformal time $d\eta \equiv dt/R(t)$. Suppose the light is emitted at $r = 0$ at time t and an observer on the earth receives it at $r = r$ at $t = t_0$. The comoving distance can be calculated if the redshift z of the light emitter is given.

$$\eta_0 - \eta = \chi_0 - \chi = \int_0^r \frac{dr}{\sqrt{1-kr^2}} = \int_t^{t_0} \frac{dt}{R(t)} = \int_R^{R_0} \frac{dR}{\dot{R}R} = \frac{1}{R_0}\int_0^z \frac{dz}{H(z)} \tag{8.66}$$

$H(z)$ can be obtained as follows: As the matter and radiation density have scale dependence $\sim R^{-3}$ and $\sim R^{-4}$, respectively (see Eq. (8.54)), the Friedmann equation can be expressed as

$$H^2 = H_0^2 \frac{8\pi G_N}{3H_0^2}\left(\rho_{m0}\frac{R_0^3}{R^3} + \rho_{r0}\frac{R_0^4}{R^4} + \rho_\Lambda\right) - \frac{k}{R^2} \tag{8.67}$$

where $\rho_\Lambda = \rho_{\text{vac}}$ is defined in Eq. (8.18). At the present time, $\Omega_{r0} \sim 10^{-4}$ and can be neglected for discussing the observed events.[17] Substituting $\rho_{c0} = 3H_0^2/(8\pi G_N)$, $R_0/R = 1 + z$, and $\rho = \rho_c\Omega$ in the above equation, we obtain

$$\begin{aligned} H(z) &= H_0\left[\Omega_\Lambda + \Omega_{m0}(1+z)^3 + (1 - \Omega_{m0} - \Omega_\Lambda)(1+z)^2\right]^{\frac{1}{2}} \\ &\simeq \begin{cases} H_0\left[1 + z\left(1 + \frac{\Omega_{m0}}{2} - \Omega_\Lambda\right) + \cdots\right] & z \ll 1 \\ H_0\sqrt{\Omega_{m0}}(1+z)^{\frac{3}{2}} & \Omega_{m0}(1+z) \gg 1. \end{cases} \end{aligned} \tag{8.68}$$

---

[17] The radiation energy cannot be neglected for $z \gtrsim 3000$, that is, for discussing events in the radiation-dominated era. But, in that case, it is much simpler to count things starting from $t = 0$ (i.e., Big Bang) rather than going backwards from present time using the redshift.

As the Hubble constant at time t is expressed as a function of z, the comoving distance can be calculated using Eq. (8.68).

$$\chi \approx \begin{cases} \dfrac{1}{R_0 H_0}\left[z - \dfrac{1}{2}(1+q_0)z^2 + \cdots\right] & z \ll 1 \\ \dfrac{2}{R_0 H_0 \sqrt{\Omega_{m0}}}\left[1 - \dfrac{1}{\sqrt{1+z}}\right] = \eta_0 - \eta(z) & 1+z \gg 1/\Omega_0 \end{cases} \tag{8.69}$$

By multiplying χ with the scale factor $R(t)$, we obtain the physical distance at time t. It is the same physical distance we are accustomed to in the stationary universe, but it is not a very usable quantity in the expanding universe. Depending on what observable we use, we define the following two distances that can be directly connected to observables:

Luminosity distance: Suppose the absolute luminosity L of an observable at a distance d_L is known and one measures its intensity (flux) F per unit area per unit time. In the Euclidean geometry, we have

$$L = 4\pi d_L^2 F \tag{8.70}$$

which defines the luminosity distance d_L. In reality, the universe is expanding. Let the light be emitted at r during time interval δt in the comoving coordinate system, and it is observed at $r = 0$ (on the earth) at time $t = t_0 \sim t_0 + \delta t_0$. The surface of a sphere with fixed r is given by $S_0 = 4\pi(R_0 r)^2$ at time $t = t_0$. On the other hand, the photon number density n_γ changes as $n_\gamma \sim T^3 \sim R^{-3}$ and the total number $N_\gamma \sim n_\gamma R^3 \sim E_\gamma/(h\nu)$ is conserved. Then

$$\frac{L\delta t}{\nu} = 4\pi R_0^2 r^2 \frac{F\delta t_0}{\nu_0}. \tag{8.71}$$

Substituting $\nu/\nu_0 = \delta t_0/\delta t = 1 + z$ and using the definition Eq. (8.70), one obtains

$$d_L = R_0 r(1+z) = R_0 S_k(\chi)(1+z) = \frac{1+z}{\sqrt{-k}/R_0} \sinh\left(\frac{\sqrt{-k}}{R_0}\int_0^z \frac{dz}{H(z)}\right) \tag{8.72}$$

One sees that the luminosity distance is directly connected to the curvature of the universe. For $\Omega_0 z \ll 1$, using Eqs. (8.66) and (8.69)

$$H_0 d_L \simeq z + \frac{1}{2}(1 - q_0)z^2 + \cdots \tag{8.73}$$

Angular Diameter Distance Suppose the transverse size $D_\perp$ of an astronomical object relative to the direction of sight is known, and its visual angle is determined to be $\delta\theta$. The angular diameter distance d_A is defined by $d_A = D_\perp/\delta\theta$. In Euclidean geometry, $d_A = d_L$. But they are different in the expanding universe. Let $D_\perp$, the actual length of an arc s at $t = t_1$ when the radius r in the comoving coordinate system, is given. As $s = r\delta\theta$ and $D_\perp$ is the length at the time when the light is emitted, one obtains $D_\perp = R(t_1)s$. Then

$$d_A = \frac{D_\perp}{\delta\theta} = \frac{R(t_1)r\delta\theta}{\delta\theta} = R(t_1)r = R_0\frac{r}{1+z} = \frac{d_L}{(1+z)^2} = \frac{R_0}{1+z}S_k(\chi) \tag{8.74}$$

For $z \ll 1$, we have

$$H_0 d_A \simeq z - \frac{3}{2}\left(1 + \frac{1}{3}q_0\right)z^2 + \cdots \tag{8.75}$$

For $1 + z \gg \Omega_0^{-1}$ and a flat universe

$$d_A = \frac{R_0}{1+z} S_k(\eta_0 - \eta) = \frac{R_0}{1+z}(\eta_0 - \eta) \overset{(8.69)}{\simeq} \frac{2}{H_0\sqrt{\Omega_{m0}}(1+z)} \tag{8.76}$$

8.5.2
Horizon

Particle Horizon The maximum radius over which a source event can influence an observer (i.e., causal) is referred to as the *particle horizon*. From the observer's point of view, it is also the limit of the distance beyond which one cannot observe events in the past. This is defined by the distance over which the light can propagate in the r direction. The proper distance is obtained by multiplying the scale factor R. Considering the light path ($ds^2 = 0$), we have

$$\begin{aligned} d_H(t) &= R(t)\int_0^r \frac{dr}{\sqrt{1-kr^2}} = R(t)\int_0^t \frac{c\,dt'}{R(t')} \\ &\underset{\overline{\overline{}}}{\overset{(8.55)}{}} \begin{cases} 2ct = cH^{-1} & : \text{radiation dominated} \\ 3c(t - t_{\rm eq}) + 2ct_{\rm eq} \simeq 3ct = 2cH^{-1} & : \text{matter dominated} \\ \frac{c}{H}(e^{Ht} - 1) & : \Lambda \text{ dominated},\ H \sim \sqrt{\frac{\Lambda}{3}} \\ & \sim H_0\sqrt{\Omega_\Lambda} \end{cases} \end{aligned} \tag{8.77}$$

where $t_{\rm eq} \simeq 50{,}000$ years (Eq. (8.135)) is the time when the universe turns from RD to MD.

Event Horizon This is defined as the distance limit beyond which one cannot observe events no matter how long one waits. It is defined by the distance over which the light emitted now reaches at infinite time.

$$\begin{aligned} d_{EH}(t) &= R(t)\int_0^{r_{\max}(t)} \frac{dr}{\sqrt{1-kr^2}} = R(t)\int_t^\infty \frac{dt'}{R(t')} \\ &= \begin{cases} \infty & : \text{radiation or matter dominated} \\ \frac{1}{H} & : \Lambda \text{ dominated} \end{cases} \end{aligned} \tag{8.78}$$

In the radiation- or matter-dominated universe, the event horizon is infinite. It means that faraway galaxies, no matter how far they are located, eventually will be visible as time passes. But in the Λ-dominated universe, the event horizon is finite.

This can be understood as follows: Consider a faraway galaxy whose receding speed is close to the light velocity. Light emitted from it reaches us, so it is within the event horizon. But in the accelerating universe, its receding speed exceeds the light velocity shortly and disappears from our visible world. As time goes, more and more galaxies disappear. This is why the event horizon is finite in the accelerating universe.

Another example of the event horizon is the black hole horizon (i.e., the Schwarzschild radius). This is realized by the strong gravity, not by the expanding universe, but the outcome is the same: the existence of a boundary beyond which one can never see through.

8.6
Genesis

The hot Big Bang cosmology means that the universe was once a hot plasma of all kinds of particles. How it was created is not known. Some people believe that the universe was created from nothing, as a quantum fluctuation. If the universe was created from nothing, there is no reason to believe the dominance of matter over antimatter. Besides, in the inflation scenario which we will discuss later, any asymmetry that might have existed in the pre-inflation era would have been wiped out.

Therefore, it seems natural to assume that the initial state of the universe had the same quantum number as that of vacuum and that equal amounts of matter and antimatter existed. However, we know that the present universe is made of matter with a small fraction of antimatter. That is, the baryon asymmetry with probably the same amount of lepton asymmetry has been created at some time of the cosmic history.[18]

8.6.1
Matter Universe

It is convenient to express the matter–antimatter asymmetry as the ratio of the matter number density to that of the photon (n_B/n_γ). It should be of the order of 1 at the temperature much higher than the baryon mass (~1 GeV) because ultrarelativistic particles behave like radiation. The number density of particle species in thermal equilibrium is expected to be about equal except for the intrinsic degrees of freedom (like spin) that the particle possesses. The ratio would decrease drastically when the temperature goes down below $\Lambda_{\rm QCD}$ because the pair annihilation process ($B + \overline{B} \leftrightarrow \gamma + \gamma$) goes much faster than its reverse process. If the cosmic expansion speed is much higher than particle reaction rates, the number of particles would be frozen unless the particle decays by itself.

We have learned that the product of the cosmic scale and the temperature RT is kept constant under normal circumstances. As the photon number density is proportional to T^3 (Eq. (8.39a)), it decreases like $\sim R^{-3}$, as the universe expands.

---

[18] For $T \gg \Lambda_{\rm QCD} \sim 200$ MeV, the baryon is separated to free quarks. Therefore, the quark number is the more correct name. The convention is historical. As to the lepton asymmetry, we are not sure if it exists because we have no means to check the abundance of neutrinos or antineutrinos.

The baryon number is also conserved, so the baryon number density decreases like $\sim R^{-3}$, too. Consequently, the ratio of the baryon number density to the photon number density n_B/n_γ is a conserved number. We know that the present photon number density (~400 cm^{-3}, see Eq. (8.41)), and that the baryon number density is approximately $n_B \sim 10^{-6}\,\mathrm{cm}^{-3}$ (see the boxed paragraph below), so we have approximately $n_B/n_\gamma \sim 10^{-9}$.

Naive estimate of the baryon density: The baryon is the building block of matter in the guise of stars and galaxies. So, let us estimate the approximate value of the baryon number density n_B based on direct observations rather than using the theory. There are about $\sim 10^{57}$ nucleons in a typical star (i.e., the Sun), $\sim 10^{11}$ stars in a typical galaxy (see Table 10.1 in Chapter 10 for details of the baryon budget). Although nonluminous baryons are more abundant in the form of IGM (intergalactic medium, mostly in the form of molecular or ionized hydrogen), we simplify our argument by neglecting them, as we are trying for a naive estimate. The number of galaxies in the visible universe is estimated to be $\sim 10^{11}$. Therefore, the total number of baryons is $\sim 10^{79}$. The size of the visible universe may be approximated by $ct_0 \sim 10^{28}$ cm $\sim$3000 Mpc. Then the baryon number density is estimated to be $n_B \sim 10^{-6}/\mathrm{cm}^3$. Dividing by $n_\gamma = 410\,\mathrm{cm}^{-3}$ (see Eq. (8.41)), it approximately reproduces the more exact baryon-to-photon number density ratio $\eta_B = n_B/n_\gamma = 6.2 \pm 0.3 \times 10^{-10}$ (see Eq. (8.132)) to be obtained in the following calculations.

In the following, we discuss a few scenarios of how to create a matter universe, that is, the asymmetry of matter–antimatter balance, quantify the above arguments, and derive a more accurate baryon-to-photon number density ratio.

8.6.2
Baryogenesis

The first scenario: Statistical fluctuation.

In the hot Big Bang cosmology, equal numbers of particles and antiparticles existed. As the universe expands, it cools down and eventually gets out of thermal equilibrium. Then the most likely explanation for the baryon asymmetry is the statistical fluctuation of the thermal equilibrium. Let us see what number we will get if the asymmetry is of statistical origin.

The annihilation rate of baryons and anti-baryons is given by $\Gamma_{N\overline{N}} \simeq n_B < v\sigma_{N\overline{N}} > \simeq n_B m_\pi^{-2}$. The annihilation ceases when its rate becomes lower than the cosmic expansion rate.

$$\Gamma_{N\overline{N}} \simeq n_B < v\sigma_{N\overline{N}} > \simeq n_B m_\pi^{-2} < H \simeq T^2/M_{\mathrm{Pl}} \tag{8.79}$$

where Eq. (8.60) was used in deriving the last equality. At a temperature much below the baryon mass energy, the baryons are nonrelativistic, and the number density is given by Eq. (8.40). Then, Eq. (8.79) means that

$$\frac{n_B}{m_\pi^2} = 2\left(\frac{m_B T}{2\pi}\right)^{\frac{3}{2}} e^{-\frac{m_B}{T}} \frac{1}{m_\pi^2} \simeq \frac{T^2}{M_{\rm Pl}} \quad \rightarrow \quad T \sim 20\,\text{MeV} \tag{8.80}$$

where the baryons were assumed to be nondegenerate and we set $\mu = 0$. Using Eq. (8.58), we can estimate that the time t (at $T \simeq 20$ MeV) is about $\sim 3 \times 10^{-3}$ s. Therefore, the residual baryon number asymmetry should be of the order

$$\frac{n_B}{n_\gamma} \sim \frac{m_\pi^2 T^2/M_{\rm Pl}}{2\frac{\zeta(3)}{2\pi^2}T^3} \sim \frac{m_\pi^2}{T M_{\rm Pl}} \sim 10^{-19} \tag{8.81}$$

This is too small compared to the observed asymmetry of $\sim 10^{-9}$. It is referred to as the *annihilation catastrophe* [673].

The second scenario: The second scenario we can think of is that the number of baryons and anti-baryons are the same but at some time they are physically separated. The difficulty of this scenario is that there is no plausible mechanism to separate matter from antimatter spatially. Besides, there is no evidence to suggest that antimatter exists in abundance somewhere in the universe. We know that the solar system consists of matter because no human beings or rockets that were sent to other planets were destroyed by pair annihilations.

By cosmic ray observations, the antimatter component is known to be less than 10^{-4} at least to the distance of the Andromeda galaxy [706, 707]. If there were antimatter galaxies somewhere in the neighborhood of matter galaxies, then we should observe gamma emissions from annihilations. Its absence is strong evidence that at least the nearby clusters of galaxies (e.g., Virgo) are matter-dominated. At larger scales there is little proof.

The third scenario: The third scenario considers that the symmetry of the universe has been destroyed by baryon-number-violating processes. However, their existence alone does not produce an excess of baryons, because an equal number of anti-baryons will also be created. Sakharov in his seminal paper [708] showed that three conditions are required to generate a baryon excess.

1) The existence of baryon number nonconserving process;
2) The existence of C and CP violation;
3) The existence of thermal nonequilibrium when the baryon-number-violating processes are active.

Condition (1) is self-explanatory. In order to clarify Condition (2), let us consider two processes in which a very heavy particle X decays to two final states characterized by the baryon number B_1, B_2 with branching ratio b and $1 - b$.

$$\frac{\Gamma(X \rightarrow B_1)}{\Gamma(X \rightarrow \text{all})} = b, \qquad \frac{\Gamma(X \rightarrow B_2)}{\Gamma(X \rightarrow \text{all})} = 1 - b \tag{8.82a}$$

Then the antiparticle $\overline{X}$ decays according to

$$\frac{\Gamma(\overline{X} \to \overline{B}_1)}{\Gamma(\overline{X} \to \text{all})} = \overline{b}, \qquad \frac{\Gamma(\overline{X} \to \overline{B}_2)}{\Gamma(\overline{X} \to \text{all})} = 1 - \overline{b} \tag{8.82b}$$

According to the CPT theorem

$$\Gamma(\overline{X} \to \text{all}) = \Gamma(X \to \text{all}) \tag{8.83}$$

Consequently, assuming that the number of X and $\overline{X}$ were equal, we obtain

$$\Delta B = (b - \overline{b})B_1 + [(1-b) - (1-\overline{b})]B_2 = (b - \overline{b})(B_1 - B_2) \tag{8.84}$$

$B_1 \neq B_2$ requires the necessity of the baryon-number-violating process and $b \neq \overline{b}$ requires CP violation (see Eq. (16.63) in [1]).

Typical examples of the baryon-number-violating reactions are X and Y gauge bosons in the GUTs given in Eq. (3.22) and/or colored Higgs (H_c) which belongs to **5** (see Eq. (3.49)). In terms of ($SU(3)$, $SU(2)$, Y) quantum numbers, they may be expressed as

$$\begin{aligned} X &= \left(\mathbf{3}, \mathbf{2}, \frac{5}{3}\right) \\ H^c &= \left(\mathbf{3}, \mathbf{2}, -\frac{2}{3}\right) \end{aligned} \tag{8.85}$$

As an example, let us consider the colored Higgs. It has decay modes with different baryon number.

$$\begin{aligned} &H^c \to \overline{qq}, \quad H^c \to q\overline{l} \\ &\overline{H^c} \to qq, \quad \overline{H^c} \to \overline{q}l \end{aligned} \tag{8.86}$$

Here, $B_1 = \overline{qq}$ may be represented by the baryon number $-2/3$, and $B_2 = q\overline{l}$ by $+1/3$.

Condition (3) follows from the following arguments: Using the CPT theorem and unitarity alone, one can derive

$$\sum_r \Gamma(\overline{B} \to \overline{r}) = \sum_r \Gamma(B \to r) = \sum_{\overline{r}} \Gamma(B \to \overline{r}) \tag{8.87}$$

The first equality follows from the CPT theorem. The second equality is due to the fact that sum over r and $\overline{r}$ go over all particle species, which are equal. Applying the CPT theorem again to the above equality, we obtain

$$\sum_r \Gamma(r \to B) = \sum_r \Gamma(r \to \overline{B}) \tag{8.88}$$

In thermal equilibrium, all the r-states exist in equal numbers, and hence the asymmetry is not produced even if conditions (1) and (2) are satisfied. Thus, the departure from thermal equilibrium in the cosmic expansion is a necessary condition for the asymmetry generation.

The SM provides the CP violation. The GUTs have a built-in mechanism to violate the baryon number conservation and can provide a natural environment. The departure from thermal equilibrium can be provided, for instance, by decays of the heavy particle H^c after decoupling at the temperature $T \ll m_H$ (delayed decay mechanism). Thus many models to produce the cosmic baryon number within the framework of the GUTs were proposed [709–712].

Results depend on the models [713, 714], but it is known that the magnitude of the CP violation in the SM model is known to be too small. Reheating temperature after the inflation may not be high enough to allow the grand unification. Besides, even if the baryon excess is produced, it will be wiped out by the sphaleron effect during the electroweak phase transition at $T \lesssim 5$ TeV (see Eq. (7.88)) unless there exists a nonzero asymmetry $\Delta(B-L)$. However, in many GUT models, including $SU(5)$ and $SO(10)$, the symmetry $[\Delta(B-L)=0]$ is respected and it is hard to produce an excess that is phenomenologically viable.

Sphaleron The discovery of the sphaleron ([636], see also Section 7.1.7) opened a new possibility to generate the baryon number during the electroweak phase transition. It was argued that the SM alone can provide the baryon number, C and CP-violating processes. However, to meet the Sakharov's third condition and further for the produced baryon asymmetry not to be erased by the sphaleron effect itself, the transition has to be strongly of the first order. One possibility is to use the boundary effect of bubbles that can be created during the supercooled phase transition. This requires the Higgs mass to be significantly lower than 80 GeV [715]. However, this possibility is denied by the precision electroweak data. The second difficulty is that the CP violation in the SM is too small to produce the observed value of the baryon number-to-entropy ratio ($n_B/s \sim 10^{-10}$). A possible cure was supplied by the supersymmetric extension of the SM [716]. The CP-violation effect could be enhanced by appealing to the extra complex phases generated by the two Higgs doublets [717]. The phase transition of the first order can be realized if the stop mass is lighter than the top mass. However, it was shown that the Higgs mass needs to be less than ~110 GeV for the phase transition to be first order [718, 719]. This possibility was rejected by the discovery of the Higgs with mass 126 GeV. In conclusion, in the framework of the SM, production of the baryon asymmetry at the time of the electroweak phase transition has many difficulties, if not impossible.

8.6.3
Leptogenesis

A currently plausible model is considered to be the leptogenesis scenario, which assumes that an excess of leptons is generated by heavy Majorana neutrino decays at the GUT era [720]. Then it is converted to an excess of the baryon number during the electroweak phase transition [721, 722].

Lepton Number Generation Generation of the lepton excess can be achieved by simply adding the right-handed heavy Majorana neutrinos to the SM.

$$\mathscr{L}_{N_R} = \sum_i \left[\overline{\nu}_{iR} i\gamma^{\mu}\partial_{\mu}\nu_{iR} + \frac{M_i}{2}\overline{\nu}^c_{iR}\nu_{iR} + h_{ij}\overline{l_{jL}}\nu_{iR}\phi + h.c. \right] \tag{8.89}$$

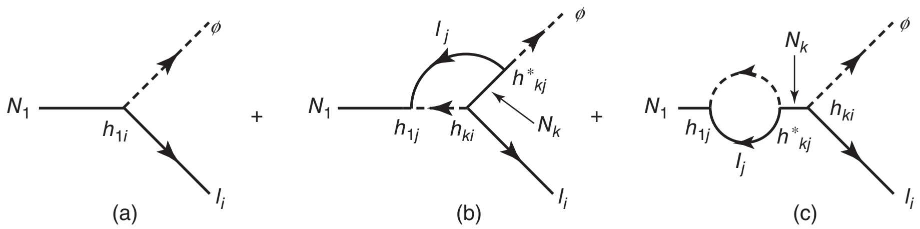


Figure 8.5 (a) Tree-level and (b) (c) one-loop-level diagrams contributing to heavy Majorana neutrino decays. The heavy Majorana neutrinos are denoted as N_k where k is the generation number. Their interference leads to lepton-number- and CP-violating decays. Complex Yukawa couplings at each vertex are displayed.

where h_{ij} are the Yukawa coupling strength, which, in general, are complex. The heavy Majorana $N_i \equiv \nu_{iR} + \nu_{iR}^c$ has lepton-number-violating decay modes.

$$N \to l_L + \overline{\phi}, \qquad N \to \overline{l_L} + \phi \tag{8.90}$$

To generate the asymmetry between the two modes, two interfering amplitudes are needed, which are supplied by the tree amplitude and higher order loop diagrams (see Figure 8.5). The magnitude of the lepton number asymmetry is calculated to be [131, 720, 723–726]

$$\begin{aligned}\varepsilon &= \frac{\Gamma(N_1 \to l_L + \overline{\phi}) - \Gamma(N_1 \to \overline{l_L} + \phi)}{\Gamma(N_1 \to l_L + \overline{\phi}) + \Gamma(N_1 \to \overline{l_L} + \phi)} \\ &\simeq \frac{1}{4\pi}\frac{1}{[hh^\dagger]_{11}}\sum_{j=2,3}\mathrm{Im}[(hh^\dagger)^2]_{1j}\left[I\left(\frac{M_j^2}{M_1^2}\right) + J\left(\frac{M_j^2}{M_1^2}\right)\right] \end{aligned} \tag{8.91}$$

$$I(x) = \sqrt{x}\left[1 + (1+x)\ln\left[\frac{x}{(1+x)}\right]\right], \quad J(x) = \frac{\sqrt{x}}{(1-x)}$$

where $[h] = \det[h_{ij}]$ and the functions $I(x)$, $J(x)$ arise from the loop integrals. For assumptions of $M_1 \ll M_3$ and h_{33} dominance among h_{ij}s, we have

$$\varepsilon \simeq \left(\frac{3}{8\pi}\right)|h_{33}|^2\left(\frac{M_1}{M_3}\right)\delta \tag{8.92}$$

where $\delta = -\sin\varphi$, with $\varphi = \arg\left([(hh^\dagger)^2]_{13}\right)$. Notice that this phase is different from the CP-violating phase which appears in the neutrino oscillation (i.e., in the PMNS matrix). The see-saw mechanism tells (see Eq. (2.18))

$$m(\nu_{3L}) \simeq \frac{m_D^2}{M_3} \simeq \frac{|h_{33}|^2 v_{\mathrm{EW}}^2}{M_3} \tag{8.93}$$

where $v_{\mathrm{EW}} \approx 250\,\mathrm{GeV}$ is the vacuum expectation value (VEV) of the SM. Substituting Eq. (8.93) in Eq. (8.92), we obtain

$$\varepsilon \simeq \left(\frac{3}{8\pi}\right)\frac{m(\nu_{3L})M_1}{v_{\mathrm{EW}}^2}\delta \simeq 1 \times 10^{-6} \times \left(\frac{M_1}{10^{10}\,\mathrm{GeV}}\right)\delta \tag{8.94}$$

For the generated asymmetry not to be erased by thermal motions, the decay has to happen after it is decoupled from the cosmic expansion (i.e., when $\Gamma(N \to l + \phi) < H$). To obtain the lepton surplus, one needs to solve the Boltzmann equation. A fitting formula to the lepton number density to specific entropy ratio is given by [727,728]

$$\begin{aligned} \frac{n_L}{s} &\simeq 0.5 \times 10^{-2} \varepsilon \frac{1}{1 + (3K)^{1.2}} \\ K &\equiv \frac{1}{2} \frac{\Gamma}{H_{T=M_{NR}}} \end{aligned} \tag{8.95}$$

L-to-B Conversion The mechanism of obtaining finite numbers of the baryon asymmetry ΔB is as follows [131,721]: Assuming that the early universe is a weakly coupled plasma, one can assign a chemical potential to each of the quark (q_{iL}, u_{iR}, d_{iR}), lepton (l_{iL}, e_{iR}), and Higgs fields. In the SM, with one Higgs doublet H and N_g generations, one has $5N_g + 1$ chemical potentials. For a noninteracting gas of massless particles, the asymmetry in the particle and antiparticle number densities can be calculated from Eq. (8.38). For a small value of μ_i/T, it gives

$$\Delta n_i = n_i - \overline{n}_i = \frac{gT^3}{6} \begin{cases} \frac{\mu_i}{T} + O\left(\left(\frac{\mu_i}{T}\right)^3\right) & : \text{fermions} \\ 2\frac{\mu_i}{T} + O\left(\left(\frac{\mu_i}{T}\right)^3\right) & : \text{bosons} \end{cases} \tag{8.96}$$

If we assume that the sphaleron is in thermal equilibrium with other particles, reactions having $\Delta(B + L) > 0$ and $\Delta(B + L) < 0$ are in equilibrium, which means

$$\sum_i \left[3\mu(q_{iL}) + \mu(l_{iL})\right] = 0 \tag{8.97}$$

where i is the generation number. When the sphaleron is active, the QCD instanton is also effective, which converts left-handed quarks to right-handed quarks, and vice versa.

$$\sum_i \left[2\mu(q_{iL}) - \mu(u_{iR}) - \mu(d_{iR})\right] = 0 \tag{8.98}$$

The hypercharge is Abelian and is not affected by the instanton. It is conserved at all temperatures in the SM. This constraint gives

$$\sum_i \left[\mu(q_{iL}) + 2\mu(u_{iR}) - \mu(d_{iR}) - \mu(l_{iL}) - \mu(e_{iR})\right] + 2\mu(H) = 0 \tag{8.99}$$

The Yukawa interactions, supplemented by the gauge interactions, yield other relations:

$$\mu(q_{iL}) - \mu(u_{jR}) + \mu(H) = 0,\ \mu(q_{iL}) - \mu(d_{jR}) - \mu(H) = 0,\ \mu(l_{iL}) - \mu(e_{jR}) - \mu(H) = 0 \tag{8.100}$$

If the Yukawa interaction is in equilibrium, we have equilibrium between the different generations and $\mu(q_{iL}) = \mu_q$, $\mu(l_{iL}) = \mu_l$. Solving (8.97)–(8.100), we can express all the chemical potentials in terms of μ_l.

$$\begin{aligned}
&\mu(q_L) = -\frac{1}{3}\mu_l, \quad \mu_H = \frac{4N_g}{3(2N_g+1)}\mu_l \\
&\mu(u_R) = \frac{2N_g-1}{3(2N_g+1)}\mu_l, \quad \mu(d_R) = -\frac{6N_g+1}{3(2N_g+1)}\mu_l, \quad \mu(e_R) = \frac{2N_g+3}{3(2N_g+1)}\mu_l
\end{aligned} \tag{8.101}$$

From Eq. (8.96), one knows that the asymmetry of the baryon and lepton-number densities is proportional to the chemical potentials, and hence it can be expressed in terms of them.

$$\Delta B = K\sum_i \left[2\mu(q_{iL}) + \mu(u_{iR}) + \mu(d_{iR})\right], \quad \Delta L = K\sum_i \left[2\mu(l_{iL}) + \mu(e_{iR})\right] \tag{8.102}$$

This yields the important relations between ΔB, ΔL, and $\Delta(B-L)$ asymmetries [729].

$$\begin{aligned}
&\Delta B = c_s\Delta(B-L), \quad \Delta L = (c_s - 1)\Delta(B-L) \\
&c_s = \frac{8N_g+4}{22N_g+13} = \frac{28}{79}
\end{aligned} \tag{8.103}$$

This means that, if $\Delta(B-L) = 0$ initially, no baryon numbers remain after the sphaleron's action.

The above relations are valid if the sphaleron is in thermal equilibrium. At high temperatures $T \gg v$, the VEV of the Higgs field is a function of the temperature $v = v(T)$, and the transition rate Eq. (7.88) due to the sphaleron does not hold any longer. The sphaleron itself does not exist at such temperatures. However, there is no apparent reason for the $B + L$ violation to be suppressed. In fact, it is argued that the B + L-violating process is in thermal equilibrium up to $T \sim 10^{12}$ GeV. Therefore, if the lepton-number generation takes place at the higher temperature, it can be effectively converted to the baryon number asymmetry.

Approximating $K \simeq 1$ in Eq. (8.95) and considering the conversion factor $c_s \sim 1/3$ to obtain ΔB from $\Delta(B-L)$, we have

$$\frac{n_B}{s} \simeq 3\times 10^{-10}\,\delta\left(\frac{M_1}{10^{10}\,\text{GeV}}\right) \tag{8.104}$$

which approximately reproduces the observed baryon number in the universe.

The beauty of the leptogenesis scenario comes from the following:

(i) The mechanism is very generic in the sense that it does not need the elaborate GUT framework.
(ii) The same Majorana neutrino that explains small masses of the light neutrinos via see-saw mechanism can be used.

The Majorana mass needed to explain the baryon asymmetry severely constrains the allowed mass range of the light neutrinos. Surprisingly, both the upper and lower limits given by the leptogenesis are consistent with the mass values constrained from beta decays and neutrino oscillations. A severe test of the model will be provided if the fine-tuned Majorana mass can predict precisely the mass range of the light neutrinos.

8.6.4
Neutrino Decoupling

In the early universe, many particle species were in thermal equilibrium. But as the universe cooled down, most of them either disappeared by pair annihilations or were decoupled from the equilibrium. At the cosmic temperature $T \simeq$ a few × MeV (or $t \sim 1$ s by Eq. (8.58)), most of the particles that were heavier than the electron had disappeared. The main thermal reactions at this stage are neutrino reactions and electron–positron pair annihilations. Because of the baryon asymmetry that we have just discussed, a small number of baryons [consisting of protons (p) and neutrons (n)] ($n_B/n_\gamma \sim 10^{-10}$) participate the thermal reactions, too.

$$\begin{aligned} &\nu_i + e^\pm \leftrightarrow \nu_i + e^\pm, \\ &\nu_e + n \leftrightarrow e^- + p, \\ &\overline{\nu}_e + p \leftrightarrow e^+ + n, \\ &\nu_i + \overline{\nu}_i \leftrightarrow e^- + e^+ \leftrightarrow \gamma + \gamma \end{aligned} \tag{8.105}$$

where the index i (=e, μ, τ) denotes the flavor of the neutrino. However, the main player is the electric neutrino, because μ and τ had already vanished. The interaction rate for each neutrino is given by

$$\Gamma \simeq n <\sigma v> \tag{8.106}$$

where n is the number density of the target particles, σ is the cross section, $v \simeq 1$ is the neutrino velocity, and $<\cdots>$ is the thermal average. As the neutrino reaction rate in Eq. (8.105) is given by $\sigma v \sim G_F^2 E^2$, where E is the center of mass energy and its average is given by the temperature, we have the relation

$$<\sigma v> \simeq G_F^2 T^2 \tag{8.107}$$

The number density of relativistic particles is given by Eq. (8.39a), which approximates as $n \simeq T^3$. Therefore, we have approximately

$$\Gamma \sim G_F^2 T^5 \tag{8.108}$$

The universe is expanding at the rate H and, when it exceeds the reaction rate, the neutrino reaction no longer happens (neutrino decoupling). As the Hubble rate is given by Eq. (8.60), we can determine the temperature of decoupling by equating

$$\Gamma = H \quad \rightarrow \quad G_F^2 T^5 \simeq \frac{T^2}{M_{\mathrm{Pl}}} \tag{8.109a}$$

$$\therefore \quad T_{\mathrm{dc}} \simeq (M_{\mathrm{Pl}} G_F^2)^{-\frac{1}{3}} \sim 1\,\mathrm{MeV} \tag{8.109b}$$

A more detailed calculation, taking into account the multichannel reactions, gives $T \simeq 1.3$ MeV [730].

It is noteworthy that the neutrino decoupling happens before the pair annihilation process ($e^- e^+ \rightarrow \gamma\gamma$). When the temperature goes down below $2m_e$ (or more precisely $T \simeq 0.17$ MeV), the electron–positron annihilation becomes one-way traffic and almost all of the pairs disappear to increase the number of photons. The energy of pair-annihilated particles is consumed to raise the photon temperature. Some of them survive to conserve the electric charge, but at this stage their contribution can be safely neglected. As the entropy in a volume V ($S = sV$) is conserved before and after the reaction, we equate them. Using Eq. (8.47a), we have

$$\left[\frac{4}{3}\frac{V}{T}\left(\rho_{e^-} + \rho_{e^+} + \rho_\gamma\right)\right]_{\text{before}} = \left[\frac{4}{3}\frac{V}{T}\rho_\gamma\right]_{\text{after}} \tag{8.110}$$

Considering $V \sim R^3$, $\rho_{e^-} + \rho_{e^+} = 2\,(7/8)\rho_\gamma$ (Eq. (8.39b)), the condition for the entropy conservation becomes

$$\begin{aligned} \frac{11}{4}T^3R^3\bigg|_{\text{before}} &= T^3R^3\bigg|_{\text{after}} \\ \therefore\quad T_{\text{before}} &= \left(\frac{4}{11}\right)^{\frac{1}{3}} T_{\text{after}} \end{aligned} \tag{8.111}$$

This is the photon temperature. At this point, the neutrino is already decoupled, and therefore is not heated by pair annihilation. Before pair annihilation, $T_\nu = T_\gamma$, but after it the annihilation energy raises the photon temperature but does not change the neutrino temperature. Therefore, after the annihilation

$$T_\nu = \left(\frac{4}{11}\right)^{\frac{1}{3}} T_\gamma = 0.71T_\gamma \tag{8.112}$$

That is, from $t = t_{\nu\,dc}$ or $T \simeq m_e$, the neutrino temperature is always lower than the photon temperature.

Neutrino temperature: After the decoupling, the neutrino is no longer in thermal equilibrium with any other particle but is streaming freely. Therefore, its temperature cannot be determined in a strict sense. However, its energy spectrum is given by that at decoupling, that is, the Planck formula, which is uniquely defined by the temperature. According to Eq. (8.32), the spectrum is just redshifted as the universe expands, keeping its spectrum shape. Thus one can speak of the neutrino temperature even if it is not in thermal equilibrium.

If the neutrino is massive, it is nonrelativistic now. But the energy spectrum is given by the Planck formula. This justifies our treatment of the neutrino temperature.

The present neutrino temperature and the number density are given by

$$\begin{aligned} T_{\nu 0} &= \left(\frac{4}{11}\right)^{1/3} T_{\gamma 0} \approx 1.92\,\text{K} \\ n_{\nu 0} &= \frac{3}{4}g_i\frac{\zeta(3)}{\pi^2}T_0^3 = \frac{3}{4}\frac{4}{11}n_{\gamma 0} = 112\,\text{cm}^{-3}/\text{flavor} \end{aligned} \tag{8.113}$$

Requiring that the neutrino energy density does not exceed the cosmic dark matter energy density, we obtain constraints for the neutrino mass.

$$\Omega_{\nu 0} = \frac{\rho_{\nu 0}}{\rho_c} = \frac{8\pi G_N}{3H_0^2} \times 3n_\nu^0\,m_\nu \le \Omega_{m0} \tag{8.114}$$

from which we obtain

$$\sum m_\nu < 90\,\Omega_m h^2\,\mathrm{eV} \simeq 12\,\mathrm{eV} \tag{8.115}$$

Large-scale structure data of the galaxy distributions can also place constraints (see arguments in Section 9.1.8). For instance, the 2dF data on power spectrum sets an upper limit on the ratio $\Omega_{\nu 0}/\Omega_{m0} < 0.13$, which corresponds to $m_\nu < 2$ eV [8]. Additional cosmological datasets bring down the limit further. An upper limit on the total neutrino mass of 0.17 eV was reported by combining several cosmological probes [120, 121, 731].

8.6.5
Big Bang Nucleosynthesis

A small number of surviving quarks hadronize at $T \lesssim \Lambda_{\mathrm{QCD}} \simeq 200$ MeV. But when $T \gg 10$ MeV, the protons and neutrons are not bound, and fly freely. When the temperature goes down below the binding energy of nuclei, nuclear formation begins. A reaction in thermal equilibrium to form a nucleus with atomic number A from the proton and neutron is expressed as

$$Zp + (A - Z)n \leftrightarrow A \tag{8.116}$$

which means the chemical potential equation $\mu_A = Z\mu_p + (A - Z)\mu_n$. Therefore

$$e^{\frac{\mu_A}{T}} = e^{\frac{Z\mu_p + (A-Z)\mu_n}{T}} \tag{8.117}$$

We define the baryon number density n_B and η_B by

$$n_B = n_n + n_p + \sum_A A n_A \equiv \eta_B n_\gamma = \eta_B \left[\frac{2\zeta(3)}{\pi^2} T^3\right] \tag{8.118}$$

and their weight ratio by

$$X_A = \frac{A n_A}{n_B}, \quad X_p = \frac{n_p}{n_B}, \quad X_n = \frac{n_n}{n_B} \tag{8.119}$$

Substituting the number density formula Eq. (8.40) for each particle into Eq. (8.119) and removing the chemical potential by using Eq. (8.117), we obtain

$$\begin{aligned} X_A &= \frac{g_A}{2^A} A^{5/2} X_p^Z X_n^{A-Z} \left[n_B \left(\frac{2\pi}{m_N T} \right)^{\frac{3}{2}} \right]^{A-1} e^{\frac{B_A}{T}} \\ B_A &= Z m_p + (A - Z) m_n - m_A \end{aligned} \tag{8.120}$$

The coefficient of the exponential function in the parenthesis $[\cdots]$ can be obtained by scaling the present n_B by $(T/T_0)^3$ to give

$$n_B \left(\frac{2\pi}{m_N T} \right)^{\frac{3}{2}} = \frac{\Omega_{b0}\rho_{c0}}{m_N} \left(\frac{T}{T_0} \right)^3 \left(\frac{2\pi}{m_N T} \right)^{\frac{3}{2}} \sim 3.74 \times 10^{-12} \Omega_{b0} h^2 \left(\frac{T}{\mathrm{MeV}} \right)^{\frac{3}{2}} \tag{8.121}$$

which is a very small number. Consequently, at $T \sim 1$ MeV, while being low enough to allow nucleosynthesis to proceed, it is immediately destroyed by the photons (photodisintegration of the deuteron). The nucleosynthesis does not grow because of the nucleon's low abundance until the temperature goes down to as low as ~0.1 MeV. Then the Maxwell–Boltzmann factor ($e^{B_A/T}$) controlling thermal abundance grows large enough to compensate for the low baryon density. In the following, we follow the process more carefully step by step.

The first stage: ($T \gg 10$ MeV, $t \ll 1$ s)

During this period, all the light particles ($m \ll 10$ MeV, including ν, e^-, e^+) are relativistic. The number density of the strongly interacting baryons is much low, that is, only one particle compared to 10^{10} leptons and photons. At this stage, the weak interaction is still active because of the high particle density and the reactions described in Eq. (8.105) are in thermal equilibrium. For the reaction between the neutrino and the baryon ($\nu_e + n \leftrightarrow e^- + p$), the following chemical potential relation holds:

$$\mu_\nu + \mu_n = \mu_e + \mu_p \tag{8.122}$$

Using Eq. (8.40), we can remove μ_n and μ_p to obtain

$$\frac{n_n}{n_p} = \frac{n_n/n_B}{n_p/n_B} = \frac{X_n}{X_p} = \exp\left[\frac{-(m_n - m_p) + (\mu_e - \mu_\nu)}{T}\right] \tag{8.123}$$

The leptonic chemical potential can be neglected. The justification for this is as follows: We postulate an electrically neutral universe which leads to

$$n(e^-) = n(e^+) + n_p \tag{8.124}$$

From the equilibrium equation $e^- + e^+ \leftrightarrow \gamma + \gamma$, we have another relation $\mu(e^-) + \mu(e^+) = 0$. Then

$$e^{[\mu(e^-)-\mu(e^+)]/T} = \frac{n(e^-)}{n(e^+)} = 1 + \frac{n_p}{n(e^+)} \simeq 1 + \eta_B \tag{8.125}$$

which leads to $\mu_e/T \sim \eta_B \simeq 0$. We can only guess the neutrino chemical potential, but there is no reason to believe that it behaves differently from that of the electron. Neglecting the lepton chemical potential, we have

$$\frac{n_n}{n_p} = \frac{X_n}{X_p} = e^{-\frac{Q}{T}}, \qquad Q = m_n - m_P = 1.293\,\text{MeV} \tag{8.126}$$

For $T \gg Q$,

$$X_n = X_p = 0.5 \tag{8.127}$$

For the deuteron, substituting $A = 2$ in Eq. (8.120), we obtain

$$X_D = 4.07 \times \left[\frac{T}{m_N}\right]^{\frac{3}{2}} \eta_B e^{\frac{2.22}{T(\text{MeV})}} \simeq 6 \times 10^{-12} \tag{8.128}$$

which can be safely neglected compared to the neutron or proton.

The second stage: ($T \simeq 1$ MeV, $t \simeq 1$ s)

The neutrinos decouple at $T \simeq 1.3$ MeV. Immediately after the neutrino decoupling, the reaction $e^-e^+ \to \gamma\gamma$ becomes one-way traffic. At this point, the neutron-to-proton ratio is given by

$$\frac{n_n}{n_p} = e^{-\frac{Q}{T}} \simeq \frac{1}{6}, \qquad X_n \simeq \frac{1}{7}, \quad X_p \simeq \frac{6}{7}, \quad X_D \simeq 10^{-12} \tag{8.129}$$

The deuteron, if it is synthesized, is photodisintegrated ($\gamma + D \to p + n$) immediately and does not increase. Thorium and lithium are even scarcer.

The third stage: ($T = 0.3 \to 0.1$ MeV, $t = 1$ s$\to$ 3 min)

At this stage, although the reaction rate is high, the number density is too small for any sizable nucleosynthesis to happen in the deuteron and heavier nucleus. During the 3 minutes of the third stage, the neutron decays freely with lifetime $\tau = 885.7 \pm 0.8$ s. At $T = 0.07$ MeV, the temperature is low enough compared to the helium binding energy ($B_{\text{He}} = 28.3$ MeV) such that $e^{B_A/T}$ in Eq. (8.120) grows to make the mass fraction of nuclear components comparable to $O(1)$. At this stage, reactions to form stable helium progress rapidly.

$$\begin{aligned} &D + D \to n + {}^3\text{He}, \qquad && D + {}^3\text{He} \to p + {}^4\text{He} \\ &D + D \to p + {}^3\text{H}, && D + {}^3\text{H} \;\to n + {}^4\text{He}. \end{aligned} \tag{8.130}$$

Heavier and stable nuclei can be formed in principle. Practically, however, many factors, such as the nonexistence of $A = 5$ stable nuclei, the fact that the number of materials is too low, and rapid growth of the Coulomb barrier due to decreasing temperature, prevent the formation of heavier elements. The time evolution of nuclear components is described in Figure 8.6.

Putting the number of neutrons and protons as n and p, respectively, the ratio n/p is ~1/7 at the time of helium formation. Then the helium weight ratio at $t = 3$ min can be obtained as

$$X_{\text{He}} = \frac{4(n/2)}{n+p} = \frac{2(n/p)}{1+n/p} \simeq \frac{2/7}{8/7} \simeq \frac{1}{4} \tag{8.131}$$

This is an important result. It means that the cosmic abundance of hydrogen and helium are given by 75% and 25%, which has been confirmed by observations.[19]

Gamow, in his original papers on the Big Bang cosmology, proposed that all the elements are also made at the time of the Big Bang [734–738]. Today, it is known that elements heavier than helium are synthesized in the stars and during supernova explosion. Elements that are synthesized in the early universe are limited to light elements. D, ^{3}He, and ^{7}Li are also synthesized but their abundance is exceedingly small. But using the above scenario, fine-tuned values with observations of such rare elements have reproduced the observational data

---

[19] The He/H ratio critically depends on the lifetime of the neutron and the expansion rate of the universe at this time. The latter is affected by energy density which in turn by the number of neutrino species n_g. Conversely constraints on n_g can be placed from measurements of the ratio and by extension from CMB spectrum (See Figure 2.1).

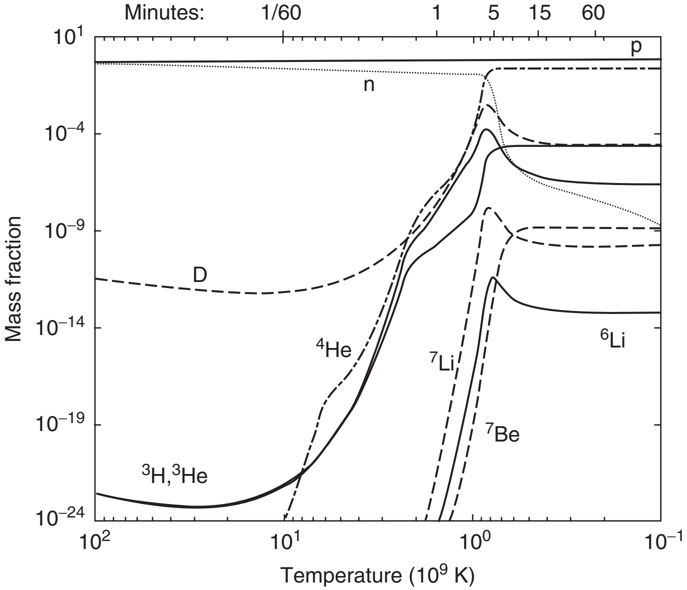


Figure 8.6 Evolution of nuclear components in the universe. The abscissa is temperature T (below) and time (top). The ordinate is mass fraction of nuclei relative to hydrogen. (Reproduced with permission of [125, 732, 733].)

quite well (see Figure 8.7). Reproduction of the cosmic light element abundance has provided one of solid foundations to support the Big Bang cosmology. The fitted values contain $\eta_B = n_B/n_\gamma$ as an adjustable parameter. The figure shows that the allowed range of η_B is very narrow. Combined with the WMAP CMB data (see Figures 9.11 and 9.13 and discussion in Section 9.2.6), one obtains [7]

$$\eta_B = \frac{n_B}{n_\gamma} = 2.75 \times 10^{-8} \times \Omega_{b0} h^2 = 6.2 \pm 0.3 \times 10^{-10}, \tag{8.132a}$$

$$\Omega_{b0} \equiv \frac{\rho_{b0}}{\rho_{c0}} = 0.0227 \pm \frac{0.0006}{h^2} = 0.044 \pm 0.004,^{20)} \tag{8.132b}$$

$$h = 0.72 \pm 0.03. \tag{8.132c}$$

The value represents the most accurate baryon number density of the present universe. Remember that the baryons are the main components of the stars, galaxies, and intergalactic medium, which occupy most of visible skies. That they occupy only 4.4% of the total cosmic energy budget (see Table 9.1) is a fact that one should remember. It also provides one of the firmest supporting evidences for the existence of dark matter, which will be discussed later.

20) The Planck data [123] combined with the WMAP data gives $\Omega_{b0} h^2 = 0.02205 \pm 0.00028$, $h = 0.6780 \pm 0.0077$.

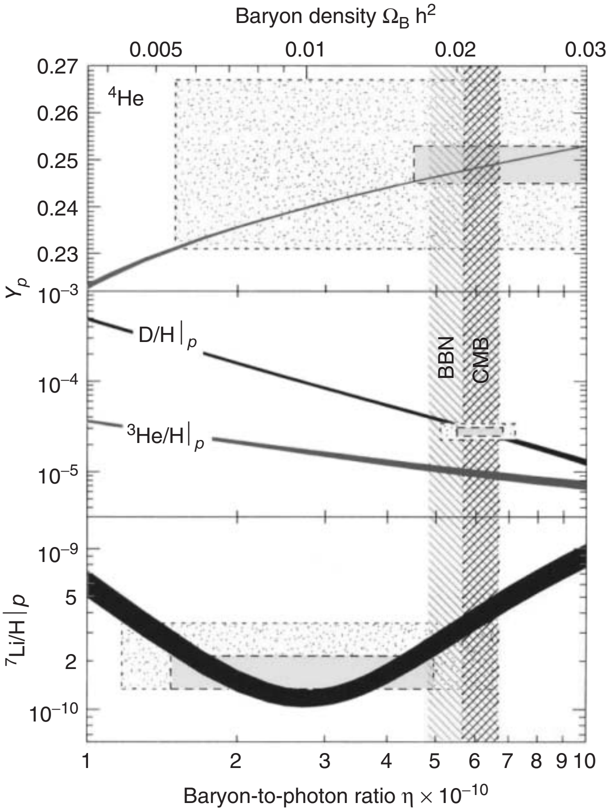


Figure 8.7 Baryon components of the universe. The unit of abscissa is $\eta_{10} \equiv n_B/n_\gamma \times 10^{10}$. Values determined from observation of light elements agree well with those determined from the nucleosynthesis [733, 739] and CMB (cosmic microwave background radiation) spectrum by WMAP [705, 740]. (Reproduced with permission of [7].)

8.7
Last Scattering

8.7.1
Radiation–Matter Equality

This is a quiet event that has happened before the much-heralded event "recombination." Nevertheless, it is an important event that has formatted the present structure of the universe. It is the time zero when the structure formation of galaxies began to grow, which will be the main subject of the next chapter.

Let us calculate the time of transition (t_{eq}) from the RD to the MD. As the radiation/matter energy density ρ_{rad}/ρ_m and the cosmic scale R are related by

Eqs.(8.54) and (8.55), they satisfy the relations

$$\begin{aligned}\rho_{\text{rad}}(t_0)R^4(t_0) &= \rho_{\text{rad}}(t_{\text{eq}})R^4(t_{\text{eq}}),\\ \rho_m(t_0)R^3(t_0) &= \rho_m(t_{\text{eq}})R^3(t_{\text{eq}}).\end{aligned} \tag{8.133}$$

Notice the matter density here ($\Omega_{m0}h^2 = 0.133 \pm 0.006$ [8]) includes the dark matter contribution and the radiation includes the neutrino as well as the photon. Consequently, by setting $\rho_{\text{rad}}(t_{\text{eq}}) = \rho_m(t_{\text{eq}})$, we obtain

$$\begin{aligned}z_{\text{eq}} + 1 &= \frac{R_0}{R(t_{\text{eq}})} = \frac{\rho_m(t_0)}{\rho_{\text{rad}}(t_0)} \overset{(8.52)}{=\!=} \frac{\rho_c\Omega_{m0}}{1.6813\,\rho_{\text{CMB}}}\\ &= 2.407 \times 10^4\Omega_{m0}h^2 \simeq 3200 \times \left(\frac{\Omega_{m0}h^2}{0.133}\right).\end{aligned} \tag{8.134}$$

The temperature at the radiation–matter equality is thus given by

$$T_{\text{eq}} = T_0(1 + z_{\text{eq}}) \simeq 0.752\,\text{eV}$$

$$t_{\text{eq}} \overset{(8.58)}{=\!=} \left(\frac{90}{32\pi^3 G_N\, g_*(T_{\text{eq}})}\right)^{\frac{1}{2}} T^{-2} = \frac{2.42}{\sqrt{g_*(T_{\text{eq}})}}\left(\frac{T}{\text{MeV}}\right)^{-2} = 49,\ 800\,\text{years} \tag{8.135}$$

where $g_*(T_{\text{eq}}) \simeq N(T \ll m_e) = 7.25$ was used. The time of matter–radiation equality is a little while before the recombination described in the next section. For the sake of later convenience, we also give the horizon at $t = t_{\text{eq}}$, which plays an important role in discussing the power spectrum of the cosmic density fluctuation.

$$d_{H\,\text{eq}} = \frac{1}{H_{\text{eq}}} \equiv \frac{1}{k_{\text{eq}}}, \qquad H_{\text{eq}} \simeq \sqrt{2}H_0\sqrt{\Omega_{m0}}(1 + z_{\text{eq}})^{\frac{3}{2}} \tag{8.136}$$

where we used Eq. (8.68) to obtain an approximate formula for $H(z)$ with correction factor of $\sqrt{2}$ because, at $t = t_{\text{eq}}$, $\rho_{\text{rad}} = \rho_m$. The wavenumber k_{eq} and the reduced wavelength $1/k_{\text{eq}} = \lambda_{\text{eq}}/2\pi$ scaled to today are given by

$$\begin{aligned}\frac{1}{k_{\text{eq}}(t_0)} &= \frac{\lambda_{\text{eq}}(t_0)}{2\pi} = \frac{R_0}{R_{\text{eq}}}\frac{1}{H_{\text{eq}}} \simeq \frac{1}{H_0\sqrt{2\Omega_{m0}}\,(1 + z_{\text{eq}})^{\frac{1}{2}}}\\ &\simeq \frac{13.67}{\Omega_{m0}h^2}\,\text{Mpc} = 102.8\left(\frac{\Omega_{m0}h^2}{0.133}\right)^{-1}\ \text{Mpc}.\end{aligned} \tag{8.137}$$

8.7.2
Recombination

When the temperature drops down to a critical value T_{rc}, the photon can no longer prevent the electrons and protons (and helium too) to form neutral hydrogen and helium atoms. The occasion is referred to as the *recombination*, although it happened for the first time. The temperature at recombination can be calculated as follows: We neglect contribution of helium for simplicity. Let electrons, protons, and hydrogen are in chemical equilibrium, that is

$$e^- + p \leftrightarrow \text{H} + \gamma \tag{8.138}$$

As the chemical potential of the photon is zero, the following equality holds:

$$\mu_e + \mu_p = \mu_{\rm H} \tag{8.139}$$

If the net electric charge of the universe is zero, we have

$$n_e = n_p, \qquad n_p + n_{\rm H} = n_B \tag{8.140}$$

where n_B is the number density of the baryon. Using the nonrelativistic number density

$$n = g\left[\frac{mT}{2\pi}\right]^{\frac{3}{2}} e^{\frac{\mu-m}{T}} \tag{8.141}$$

the relation among chemical potentials Eq. (8.139) can be converted to that of number densities. As a result, we obtain

$$n_H = n_e n_p \left[\frac{g_H}{g_e g_p}\right]\left[\frac{m_H}{m_p}\frac{2\pi}{m_e T}\right]^{\frac{3}{2}} e^{\frac{B}{T}} \tag{8.142}$$

$$B = m_e + m_p - m_H = 13.6\,\text{eV} \tag{8.143}$$

Introducing

$$\begin{aligned} X_e &\equiv \frac{n_e}{n_B} = \frac{n_p}{n_B}, \quad n_B = \eta_B n_\gamma, \quad \eta_B = 2.75\times 10^{-8}(\Omega_b h^2) \\ n_\gamma &= g_\gamma \frac{\zeta(3)}{\pi^2}T^3, \qquad \zeta(3) \approx 1.202 \end{aligned} \tag{8.144}$$

and setting $g_H = 4$, $g_p = g_e = 2$, we obtain

$$\frac{n_H n_B}{n_p n_e} = \frac{1-X_e}{X_e^2} = n_B\left[\frac{2\pi}{m_e T}\right]^{\frac{3}{2}} e^{\frac{B}{T}} = \frac{4\sqrt{2}\zeta(3)}{\sqrt{\pi}}\eta_B\left[\frac{T}{m_e}\right]^{\frac{3}{2}} e^{\frac{B}{T}} \tag{8.145}$$

The above equation is referred to as the *Saha equation* in thermal equilibrium for ion recombination. For quantitative calculation, one has to include the helium contribution. Putting $T = T_0(1+z_{\rm rec}) = 2.73(1+z_{\rm rec})$, and defining the recombination temperature $T_{\rm rec}$ by that when the value of $X_e = 0.1$, we obtain

$$1 + z_{\rm rec} \simeq 1300, \qquad T_{\rm rec} \simeq 0.305\,\text{eV} = 3540\,K \tag{8.146}$$

Naively, one might think that recombination occurs around $T = 13.6$ eV, which is the hydrogen bound state energy. However, the number of photons is so large ($\eta_B = n_B/n_\gamma = 2.7\times 10^{-8}\Omega_b h^2$) and Bose–Einstein distribution of the photon produces a long tail at the high-temperature side. The photon can efficiently ionize hydrogen down to $T \sim 0.3$ eV.

When the temperature goes below $T = 0.305$ eV, X_e decreases rapidly, and stabilizes at $T \simeq 0.256$ eV= 2970 K to a value $X_\infty \simeq 10^{-4} \sim 10^{-5}$, which defines the decoupling time $t_{\rm dc}$. Precise values are obtained by fitting the spatially flat ΛCDM model (see Section 8.3.3) to WMAP data [705, 740, 741]. We have [7]

$$z_{\mathrm{dc}} = 1091 \pm 1, \tag{8.147a}$$

$$T_{\mathrm{dc}} = 0.256\,\mathrm{eV} = 2970 \pm 3\,\mathrm{K}, \tag{8.147b}$$

$$t_{\mathrm{dc}} = 379,000 \pm 5000\,\mathrm{year}. \tag{8.147c}$$

The decoupling happened after the recombination. Their timing is slightly different but they are used interchangeably in many occasions. The mean free path of the photons is determined by residual ionizations. As the dominant mechanism for scattering the photon is the Thomson scattering and its mean free path exceeds the size of the observable universe after the decoupling, the photon travels freely and reaches us. The oldest light we observe today comes directly from the surface of a sphere at $z = 1090$, which happened at $t \simeq 400\ 000$ years and is referred to as the *sphere of last scattering*.

After the decoupling, the number of photons in a comoving volume is maintained (freeze out). The photon spectrum is only redshifted, keeping its spectral shape, which is observed today as the CMB. This is the earliest visible record of the hot Big Bang universe. Note that, if we could observe the relic neutrinos, we can trace the record back to 1 s after the Big Bang.

8.7.3
Dark Age

Hydrogen atoms are reionized by the first born star lights around 3 billion years.[21] While the universe is transparent to the decoupled photons, they are out of the visible range. Neutral hydrogens are opaque to visible light because photons are absorbed to excite electrons to higher levels. This epoch, therefore, is referred to as the *dark age*. When the stars are formed, their light ionizes hydrogen and later helium, too. Then the sky is illuminated by sparkles of stars. This is referred to as the re-ionization.

Lights from quasars, especially those known as the Lyman α series, which is emitted when excited electrons fall to the lowest level of the hydrogen, is a good indicator of the end of the dark age. If there is a cloud of neutral hydrogen in the path of light from the quasars, they exhibit absorption lines at the corresponding redshift z. Therefore, the absorption spectra of neutral hydrogen, known as the *Lyman α forest*, are powerful tools to investigate the distribution of IGM (i.e., hydrogens) and serve to clarify the structure of the universe at the small scale. Quasars of up to $z = 7$ were observed. So the reionization happened sometime at $7 \lesssim z$. Another estimate comes from WMAP observation of the CMB anisotropy (see Section 9.2.6), which indicates that it happened at $z \simeq 11$. Combining the two analyses together, we conclude that the reionization happened sometime at $7 < z \lesssim 11$.

---

[21] The present best value of redshift at reionization is $z = 11.0 \pm 0.4$ [8].

8.8
Inflation

8.8.1
Slow Rolling and Reheating

The Big Bang cosmology has been extremely successful in explaining the observed universe. However, it has several conceptual difficulties. We list five of them below.

(i) The horizon problem, (ii) the flatness problem, (iii) the monopole problem, (iv) the cosmological constant problem, and (v) the coincidence problem.

The inflation model was devised to solve the problems (i)–(iii). The problem (iv) and (v) will be discussed in Section 11.2. The model proposes that the universe has gone through a rapid exponential expansion in the early epoch of the universe. Such rapid expansion is possible if the cosmological constant term is the dominant contribution to the cosmic energy density.

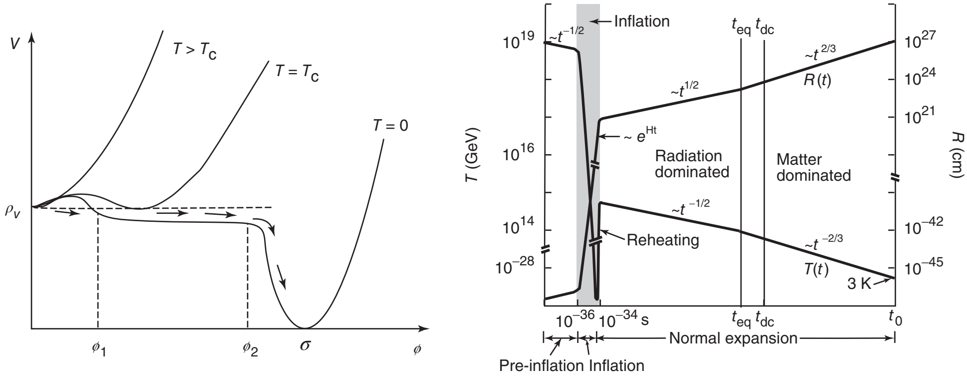
Looking at the scale growth for Λ-dominated universe in Eq. (8.55), the time dependence can be written as

$$R(t) = R(t_i)e^{H_\Lambda(t-t_i)}, \qquad H_\Lambda = \sqrt{\frac{\Lambda}{3}} \overset{(8.18)}{=\!=} \sqrt{\frac{8\pi G_N \rho_v}{3}} \tag{8.148}$$

where t_i is the time when the inflation begins. The vacuum energy can be produced when a scalar field which permeates the universe (referred to as the *inflaton*) breaks the grand unification symmetry spontaneously and induces a cosmic phase transition. Let the temperature be T_c at which the phase transition occurs. At the temperature $T \gg T_c$, the GUT symmetry is not broken, and the vacuum is at $\phi = 0$, where ϕ is the inflaton field. There are many inflaton models [742, 743] which do not necessarily coincide with what is required for the GUT phase transition from the perspective of particle physics as discussed in Chapter 3. In this case, the inflaton is considered as a singlet in the GUT gauge symmetry group, which interacts only weakly with ordinary particles in the SM [744].

Let us consider a case in which the inflaton has the potential depicted in Figure 8.8a. Depending on details of the model, symmetry breaking occurs through either the first-order phase transition or the second-order phase transition. In the former, the field tunnels through a potential barrier. The transition is violent, like boiling water releasing a fixed amount of latent heat. For the assumed potential in Figure 8.8a, the first stage of the symmetry breakdown is the path from $\phi = 0$ to ϕ_1, which is a first-order phase transition. During $\phi = \phi_1 \sim \phi_2$, the vacuum point moves slowly. The potential energy stays almost a constant with very small kinetic energy. It is the inflation epoch. For $\phi > \phi_2$, the vacuum falls down quickly to $\phi = \sigma$, and begins to oscillate (generation of inflaton particles which decay subsequently). Eventually, the vacuum settles at σ. Until the inflaton settles at $\phi = \sigma$, the vacuum has energy $\sim\rho_v$. The equation of motion for the inflaton field, which takes into account the cosmic expansion, can be obtained from the action

$$\int d^3x\, \mathscr{L}_\phi \to \int d^3x\, R^3(t)\mathscr{L}_\phi = \int d^3x\, R^3 \left(\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi)\right) \tag{8.149}$$



(a) (b)

Figure 8.8 (a) Slow roll-over inflation model. For its behavior see text. (b) Time variation of the temperature and cosmic scale in the inflation scenario. After passing the pre-inflation era ($t \lesssim 10^{-36}$ s, $R \sim t^{1/2}$), the universe expands by $e^{100} \sim 10^{43}$ during the inflation epoch ($10^{-36} \lesssim t \lesssim 10^{-34}$ s, $R \sim e^{Ht}$). After the inflation is over, reheating of the universe commences and the expansion rate of $R \sim t^{1/2}$ resumes. At $t = t_{\text{eq}} \simeq t_{\text{dc}} \sim 10^{12}$ s, the universe becomes matter-dominated and proceeds with expansion rate $R \sim T^{-1} \sim t^{2/3}$. Scale of the drawing is arbitrary.

which results in the equation of motion

$$\ddot{\phi} + \left\{3H(t) + \Gamma_\phi\right\}\dot{\phi} + \frac{\partial V(\phi)}{\partial \phi} = 0 \tag{8.150}$$

If the ϕ is considered as the position of a particle, this is the familiar equation of motion with friction. The Hubble expansion produces a dragging force for otherwise oscillating field. Γ_ϕ does not follow from the Lagrangian Eq. (8.149) but is included here to illustrate the post-inflation interaction. It is a damping factor that is induced by the inflaton interacting with other particles, which converts the inflaton to other particles. Until quanta of the inflaton are produced by oscillation, the inflaton behaves like a classical field and this term can be neglected. This is the equation of a damped oscillation. Assuming the potential shape given in Figure 8.8a, its behavior can be described qualitatively as follows: The energy density of the inflaton can be calculated from the energy–momentum tensor

$$T^\alpha_\beta = g^{\alpha\nu}\partial_\nu\phi\partial_\beta\phi - g^\alpha_\beta\left[\frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi + V(\phi)\right] = g^\alpha_0 g^0_\beta\dot{\phi}^2 - g^\alpha_\beta\left[\frac{1}{2}\dot{\phi}^2 - V(\phi)\right] \tag{8.151a}$$

$$\rho_\phi = T^0_0 = \frac{1}{2}\dot{\phi}^2 + V(\phi) \tag{8.151b}$$

$$P = T^i_i\,(\text{no sum on } i) = \frac{1}{2}\dot{\phi}^2 - V(\phi) \tag{8.151c}$$

where we have assumed a homogeneous field for which $\nabla_i\phi = 0$. For $\dot{\phi} \simeq 0$, we have $P \simeq -\rho$, which satisfies the vacuum energy condition, and the universe expands exponentially.

1) *Slow rolling*: The cosmic expansion ($3H$ in Eq. (8.150)) provides an effective friction and the first term can be neglected. The inflaton slides down slowly because of friction. The interaction of the inflaton with other particles are not turned on yet.[22)]
   If the slope of the potential between ϕ_1 and ϕ_2 is gentle and long, the time τ it takes to reach ϕ_2 (i.e., the duration that the inflation lasts) is long ($\tau \gg H^{-1}$). Incidentally, a small bump between $\phi = 0$ and ϕ_1, if it exists, induces the tunneling effect, and a phase transition of the first order occurs, generating latent heat. Bubbles of new vacuum are formed in the space of false vacuum. But here, we neglect its effect. The first-order transition used to be a common feature in early models of the inflation,
2) *Reheating*: During the inflation period, the universe is in a supercooled state. After passing the point ϕ_2, the inflaton falls down to the bottom of the potential rapidly and begins to oscillate around the potential minimum. The oscillation is fast compared to the cosmic expansion rate ($\omega \gg H, \omega^2 \sim \partial^2 V/\partial\phi^2$). Commencement of the oscillation means, quantum mechanically, the appearance of high-energy inflaton particles with mass $m \sim \omega \sim 10^{15}$ GeV, which, in turn, decays to particle–antiparticle pairs. At this point, $\Gamma_\phi \gg 3H$. In other words, the vacuum energy has been converted to high-energy particles and the inflation stops. This is a damped oscillation but the damping time is much shorter than the expansion time. When the particles are thermalized, the universe which was supercooled by the exponential expansion is reheated to the GUT temperature ($T \sim 10^{15}$ GeV) and goes back to the standard scenario of the hot Big Bang cosmology (see Figure 8.8b).[23)]

The slow-rolling condition is expressed as

$$\frac{1}{2}\dot{\phi}^2 \ll V, \qquad \ddot{\phi} \ll V', \qquad V' \equiv \frac{\partial V}{\partial \phi} \tag{8.152}$$

Consistency with Eq. (8.150) requires

$$\dot{\phi} = -\frac{V'}{3H} \tag{8.153}$$

Using $H^2 \simeq (8\pi G_N/3)V$ for $\dot{\phi}^2 << V$, the slow-rolling condition can be converted to conditions

$$\epsilon \equiv \frac{\tilde{M}_{\text{Pl}}^2}{2}\left(\frac{V'}{V}\right)^2 \approx -\frac{\dot{H}}{H^2} \ll 1, \qquad \eta \equiv \tilde{M}_{\text{Pl}}^2\left(\frac{V''}{V}\right) \ll 1 \tag{8.154}$$

[22)] At this stage, $V \approx$ constant, which means the inflaton mass is given by ($m^2 = \partial^2 V(\phi)/\partial\phi^2 \approx 0$). Spatial dependence of ϕ (Fourier component $k \sim 1/\lambda$) is diminished by the rapid expansion and redshifted, which results in $k \to 0$. Therefore, the inflaton behaves like a Bose-condensed fluid of almost massless particles with $p \simeq 0$.

[23)] Note that the reheating temperature could have been much lower. For instance, there is an argument in the supersymmetric models that overabundant relic gravitinos would have been produced if the reheating temperature T_R exceeds 10^{12} GeV [745, 746].

where the reduced Planck mass is defined by $\tilde{M}_{\rm Pl}^2 = 1/8\pi G_N$. ϵ and η are defined as indicators for the slow rolling. Notice that, as

$$\frac{\ddot{R}}{R} = \dot{H} + H^2 = (1-\epsilon)H^2 \tag{8.155}$$

inflation can be attained only if $\epsilon < 1$. As soon as this condition fails, inflation ends. These conditions determine the duration of slow roll and ensure that $P_\phi \simeq -\rho_\phi$, which is necessary for exponential expansion. The number of e-folds of the scale growth as ϕ rolls from $\phi_1 \to \phi_2$ is given by

$$N(\phi_1 \to \phi_2) = \ln\left(\frac{R_2}{R_1}\right) = \int_{t_1}^{t_2} dt\, H = \int_{\phi_1}^{\phi_2} d\phi\, \frac{H}{\dot{\phi}} = -\frac{1}{\tilde{M}_{\rm Pl}^2}\int_{\phi_1}^{\phi_2} d\phi\, \frac{V}{V'} . \tag{8.156}$$

For a polynomial potential, $V/V' \sim \phi$ and $\int(V/V') \sim \phi^2 \sim V/V''$. Thus the slow-roll condition Eq. (8.154) ensures a large N.

Setting the VEV scale of the inflaton at the GUT scale ($\sigma \sim 10^{15}$ GeV), we can estimate the Hubble time at the time of inflation. As the energy density is proportional to the fourth power of the temperature ($\rho_v \sim \sigma^4$), we have

$$H_\Lambda = \sqrt{\frac{8\pi}{3} G_N \rho_v} = \frac{1}{\sqrt{3}}\frac{\sigma^2}{\tilde{M}_{\rm Pl}} \sim 10^{11}\ \text{GeV} \sim 10^{36}\ \text{s}^{-1} \tag{8.157}$$

That is, the inflation begins at $t_i \sim 10^{-36}$ s. Let us suppose that we set the slow-rolling period $\tau = 100\, H_\Lambda^{-1} \sim 10^{-34}$ s. The length of the time looks short, but during the time span (10^{-36} to 10^{-34} s), the universe expands 100 e-fold.

$$e^{H_\Lambda \tau} \sim e^{100} \sim 3 \times 10^{43} \tag{8.158}$$

This is a big number. Consider a quark which is smaller than 10^{-16} cm. By multiplying this number, the size exceeds that of the whole universe. The condition for not contradicting the observations is generally satisfied by the number $\sim E_{\rm GUT}/T_0 \sim 10^{26} \sim e^{60}$ (see Eq. (8.164) [125]).[24] We next describe problems that we mentioned in the beginning of this section and how they are solved by the inflation model.

8.8.2
Horizon Problem

To simplify arguments, we define the horizon by $d_H(t) = ct$, which grows faster than the scale of the matter-dominated universe (grows like $\sim t^{2/3}$, see Figure 8.9a). The present horizon is $d_H(t_0) = ct_0 \sim 10^{28}$ cm, which gives the maximum observable distance. Suppose the present horizon was $d_{H0}(t)$ at time t; then it is given by

$$d_{H0}(t) = \frac{R(t)}{R(t_0)} d_H(t_0) = \frac{R(t)}{R(t_0)} ct_0 \tag{8.159}$$

Consequently

$$\frac{d_{H0}(t)}{d_H(t)} = \frac{R(t)}{R(t_0)}\left(\frac{t_0}{t}\right) \tag{8.160}$$

---

[24] Note that this estimate of the required number of e-folds assumes GUT scale reheating. For lower reheating temperatures, fewer e-folds are sufficient.

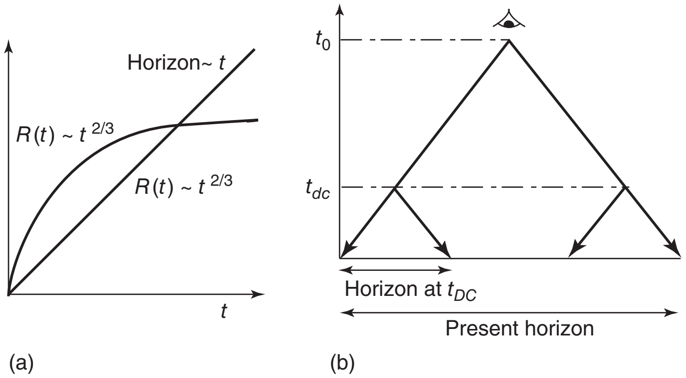


Figure 8.9 (a) The horizon grows faster than the scale of matter-dominated universe. (b) The present horizon is hundred times larger than the horizon at the decoupling.

We approximate again for simplicity that the time when the CMB was decoupled almost equals the matter-radiation equality time $t_{\mathrm{dc}} \simeq t_{\mathrm{eq}} \sim 10^{10}$ s. Then, according to Eq. (8.55)

$$\begin{aligned}\frac{d_{H0}(t)}{d_H(t)} &= \frac{R(t)}{R(t_{\mathrm{dc}})}\frac{R(t_{\mathrm{dc}})}{R(t_0)}\frac{t_0}{t} = \left(\frac{t}{t_{\mathrm{dc}}}\right)^{\frac{1}{2}}\left(\frac{t_{\mathrm{dc}}}{t_0}\right)^{\frac{2}{3}}\left(\frac{t_0}{t}\right)\\ &= \left(\frac{t_0}{t_{\mathrm{dc}}}\right)^{-\frac{1}{6}}\left(\frac{t_0}{t}\right)^{\frac{1}{2}} \approx 10^{-1}\left(\frac{t_0}{t}\right)^{\frac{1}{2}}.\end{aligned} \tag{8.161}$$

Therefore, the horizon at the time of decoupling $d_H(t_{\mathrm{dc}})$ is given by

$$d_H(t_{\mathrm{dc}}) \simeq \frac{d_{H_0}(t_{\mathrm{dc}})}{100} \tag{8.162}$$

which is 100 times smaller than the present horizon at that time (Figure 8.9b). To a very good approximation, the observed CMB is represented by the black body radiation formula with $T = 2.728 \pm 0.002$ K. The uniformity of CMB is measured to be very good; its fluctuation is within $\sim 10^{-5}$ over the whole sky. It means that the CMB is well mixed together, that is, the whole sky is within the horizon. But Eq. (8.162) shows that the present horizon is 100 times larger than the horizon at the time of decoupling. The horizon at t_{dc} observed now should be contained within an angle of visibility ~1.8°. Therefore, the uniformity of the CMB over the whole sky cannot be explained. This is the horizon problem.

Solution: Putting t_f as the time when the inflation ceased,

$$\begin{aligned}\frac{d_{H0}(t)}{d_H(t)} &= \frac{R(t)R(t_f)R(t_{\mathrm{dc}})t_0}{R(t_f)R(t_{\mathrm{dc}})R(t_0)t} = \frac{e^{H_\Lambda t}}{e^{H_\Lambda t_f}}\left(\frac{t_f}{t_{\mathrm{dc}}}\right)^{\frac{1}{2}}\left(\frac{t_{\mathrm{dc}}}{t_0}\right)^{\frac{2}{3}}\left(\frac{t_0}{t}\right)\\ &\sim 10^{-1}\left(\frac{t_0}{t}\right)^{\frac{1}{2}}\left(\frac{t_f}{t}\right)^{\frac{1}{2}} e^{H_\Lambda(t-t_f)}\end{aligned} \tag{8.163}$$

As the horizon problem exists for any phenomenon that has its origin after the GUT epoch, we require that Eq. (8.163) should be much smaller than 1 for $t = t_i \simeq 1/H_\Lambda \simeq 10^{-36}$ s. Substituting $t_f/t_i \simeq (t_f - t_i)/t_i = N$ into Eq. (8.163) gives

$$\frac{d_{H0}(t_i)}{d_H(t_i)} \simeq 10^{-1} \left(\frac{10^{17}}{10^{-36}} \right)^{\frac{1}{2}} N^{\frac{1}{2}} e^{-N} \simeq 10^{26} N^{\frac{1}{2}} e^{-N} \ll 1 \tag{8.164}$$

If $N \gtrsim 63$, we obtain the horizon before the inflation, which is much bigger than the present one. Therefore, the condition Eq. (8.159), which assumes 100 e-fold expansion by the inflation, is more than enough. Thus the horizon at t_{dc} is also much bigger than the present horizon, which solves the horizon problem.

8.8.3
Flatness Problem

The present cosmic energy density measured in units of critical density is

$$\Omega_0 = \Omega_{m0} + \Omega_{r0} + \Omega_\Lambda = 1 - \Omega_k = 1.006 \pm 0.006 \tag{8.165}$$

Theoretically, a nonzero curvature requires a very fine tuning. Suppose the present value is different from zero, that is, $\Omega_0 - 1 \sim O(1) \neq 0$.

$$\Omega(t) - 1 = \frac{\rho(t) - \rho_c(t)}{\rho_c(t)} = \frac{k}{H^2(t)R^2(t)} = \frac{k}{\dot{R}^2} \tag{8.166}$$

According to Eq. (8.55), $\dot{R} \sim t^{-1/2}$ or $t^{-1/3}$ depending on whether the times is the radiation- or matter-dominated era. Therefore

$$\frac{\Omega(t) - 1}{\Omega_0 - 1} = \frac{\dot{R}^2(t_0)\dot{R}^2(t_{\mathrm{eq}})}{\dot{R}^2(t_{\mathrm{eq}})\dot{R}^2(t)} = \left(\frac{t_{\mathrm{eq}}}{t_0} \right)^{\frac{2}{3}} \left(\frac{t}{t_{\mathrm{eq}}} \right) \tag{8.167}$$

Then

$$\begin{aligned} \Omega(t_{\mathrm{eq}}) - 1 &= 2 \times 10^{-4}(\Omega_0 - 1) \sim 10^{-4} \\ \Omega(t_{\mathrm{GUT}}) - 1 &\simeq 10^{-50}(\Omega_0 - 1) \simeq 10^{-50} \\ t_{\mathrm{GUT}} &= M_{\mathrm{GUT}}^{-1} \sim 10^{-36}\ \mathrm{s}. \end{aligned} \tag{8.168}$$

That is, the value of Ω approaches infinitely close to 1 as time goes back. In reverse logic, suppose $\Omega - 1 \sim O(1)$ in the Planck era ($t_{\mathrm{Planck}} = G_N^{1/2} \sim 10^{-44}$ s) and $k > 0$, the universe will go into recontraction instantly ($t < 10^{-43}$ s). If $k < 0$, it would have reached the present temperature $T \sim 2.7$ K in less than 10^{-11} s [125]. Such a fine tuning is unnatural and we are led to think that $\Omega = 1$ exactly from the beginning or $\Omega \to 1$ for some reason. This is the flatness problem.

Solution: Rewriting Eq. (8.166), we have

$$\Omega = \left[1 - \frac{k/R^2}{8\pi G_N \rho/3} \right]^{-1} \tag{8.169}$$

If inflation existed, $1/R^2 \sim \exp(-2H_\Lambda \tau) \to 0$, and $\Omega \to 1$ to an infinite degree of precision.

8.8.4
Monopole Problem

In the early universe, at the era $T \sim T_{\rm GUT}$, we believe that the grand unification of the forces, for instance $SU(5)$ GUT, was realized. As the temperature goes down, the symmetry is broken to separate into $SU(3) \times SU(2) \times U(1)$. It is known that topological defects known as the *t'Hooft-Polyakov's monopoles* are generated when a simply connected compact Lie group like $SU(5)$ is broken to $U(1)$ [626, 627]. Their mass is huge $\sim M_{\rm GUT}/\alpha_{\rm GUT} \sim 10^{-8}$ g. One cannot estimate how many of them are produced, but can assume them to be more than one in the horizon at the time of monopole generation [747, 748].

The horizon size at GUT is $\sim ct_{\rm GUT} = 3 \times 10^{10}$ cm s$^{-1} \times 10^{-34}$ s$\simeq 3 \times 10^{-24}$ cm. Presently, it is stretched to

$$\begin{aligned} d_H(t_{\rm GUT})\big|_{t=0} &= \frac{R(t_0)}{R(t_{\rm GUT})} ct_{\rm GUT} = \frac{T_{\rm GUT}}{T_0} ct_{\rm GUT} \\ &= \frac{10^{15}\,{\rm GeV}}{2.7\,K \times 8.6 \times 10^{-5} \times 10^{-9}\,{\rm GeV}} \times 3 \times 10^{-24}\,{\rm cm} \\ &\simeq 10^4\,{\rm cm}^2 . \end{aligned} \tag{8.170}$$

Since $M_{\rm monopole} \simeq M_{\rm GUT}/\alpha \sim 10^{17}$ GeV [626, 627], the present monopole energy density is at least

$$\Omega_{\rm Mono} = \frac{\rho_{\rm Mono}}{\rho_c} = \frac{\dfrac{10^{17}\,{\rm GeV}}{(10^4)^3\,{\rm cm}^3}}{1.05 \times h^2\, 10^{-5}\,{\rm GeV/cm}^{-3}} \simeq 10^{10}. \tag{8.171}$$

As the monopole is ultraheavy, it is hard to think of a mechanism to extinguish it. That is, the monopole should exist in abundance theoretically, but is not observed. This is the monopole problem. If the inflation existed, the number density of the monopole can be reduced exponentially to nil, consistent with the observation. Interestingly, the inflation was originally proposed to solve the monopole problem [747, 749].

If the universe goes through the inflation phase, both the number density of any particle species and the cosmic curvature essentially vanish. That is, the difference of the initial condition disappears almost completely. If the cosmic size before the inflation was $\sim ct_i \sim 10^{-25}$ cm, it will expand to 10^{18} cm after the inflation. If we go back to the GUT era, the present observed universe would have shrunk to

$$d_{H0}(t_{\rm GUT}) \simeq \frac{T_0}{T_{\rm GUT}} d_{H0}(t_0) \tag{8.172}$$

Substituting $d_{H0}(t_0) = ct_0 \sim 10^{28}$ cm, $T_0 \sim 3 \times 10^{-13}$ GeV, $T_{\rm GUT} \sim 10^{16}$ GeV, the size is mear 1 cm. We are observing only a small fraction of inflated universe.

9

Cosmology II: Structure Formation

9.1

Galaxy Distribution

9.1.1

Introduction

According to the cosmological principle, matter distribution should be uniform and isotropic. Actually, stars are grouped to make galaxies and the galaxies, in turn, form clusters of galaxies. Figure 9.1 shows a survey map of the galaxies within a circle of ∼600 Mpc, slice of 7.5° measured by the Sloan Digital Sky Survey (SDSS) group [750].

It is obvious that the galaxy distribution has a characteristic structure, which is referred to as the *large-scale structure*. The structured distribution of galaxies means deviation from uniformity which we have assumed so far. In the Friedmann equation, the energy density is a function of time but was assumed to be spatially flat. Now this restriction has to be relaxed.

What is plotted on the map? Brightness of stars is graded in "apparent magnitude m", where historically the brightest star has $m = 0$ and the faintest star has $m = 6$, as was defined by the Greek astronomer Hipparchus. In today's definition, a difference of 5 in the apparent magnitude corresponds to a factor 100 in brightness. Let the measured light fluxes and the apparent magnitudes of two stars be F, F_0 and m, $m_0 = 0$, where F_0 is used as a reference. Then

$$\frac{F}{F_0} = 100^{\frac{m_0-m}{5}} \tag{9.1}$$

For the reference star with $m_0 = 0$, Vega in the constellation Lyra had been used earlier, but now the standard flux is given by

$$F_0 = 2.52 \times 10^{-8} \text{watts m}^{-2} \tag{9.2}$$

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

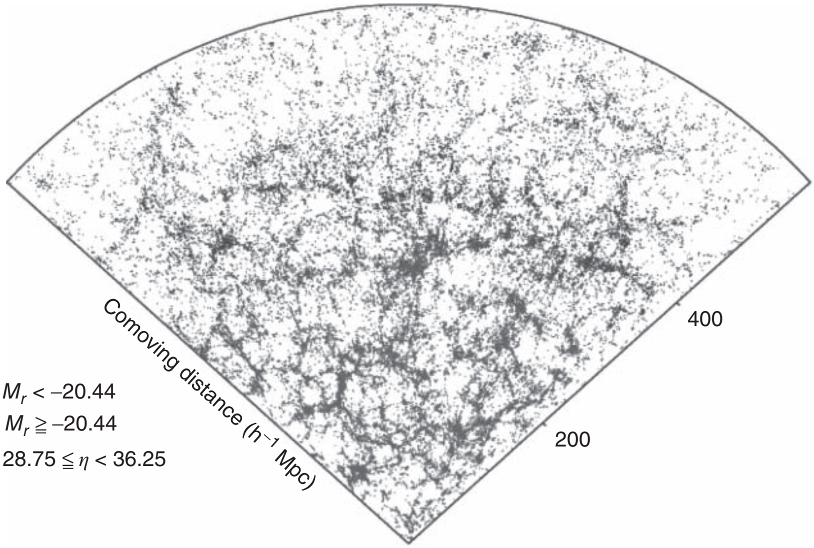


Figure 9.1 Distribution of galaxies with $14.5 \leq m_r \leq 17.5$ in 7.5° thick slice mapped by SDSS. The radial coordinate is the comoving distance (h^{-1}Mpc), and the angular coordinate is the SDSS survey longitude k. m_r and M_r are the apparent and absolute magnitudes (see boxed paragraph in the text). (Reproduced with permission of [750].)

Besides, a standard table is prepared for each color filter and is used for observations. Using the definition Eq. (9.1), the apparent magnitude is expressed as

$$m = -2.5 \log_{10} \frac{F}{F_0} \tag{9.3}$$

Astronomers attach also an absolute magnitude M to each star. It is defined as the apparent magnitude it would have if it were located at a distance of $d = 10$ pc. Knowing the apparent magnitude m, it is a calculable quantity. Then we have

$$m - M = -2.5 \left(\log_{10} \frac{F}{F_0} - \log_{10} \frac{F_{10}}{F_0} \right) = -2.5 \log_{10} \frac{F}{F_{10}} \tag{9.4}$$

The fluxes F, F_{10} are related to the star's luminosity by

$$F = \frac{L}{4\pi d_L^2}, \qquad F_{10} = \frac{L}{4\pi (10\,\mathrm{pc})^2} \tag{9.5}$$

Here, d_L is the luminosity distance defined by Eq. (8.70) and (8.72). For the second equality, we assumed that at 10 pc the red-shift is so small that the physical distance and the luminosity distance are identical. Then, Eq. (9.4) can be rewritten as

$$m - M = 5 \log_{10} \frac{d_L}{10\,\mathrm{pc}} = 5 \log_{10} \left(\frac{d_L}{\mathrm{Mpc}} \right) + 25. \tag{9.6}$$

The formula applies to the flux integrated over all wavelength region.

$$L = \int_0^\infty L(\lambda) d\lambda, \quad F = \int_0^\infty F(\lambda) d\lambda \tag{9.7}$$

When a color filter with a limited wavelength range is used, a K-correction which is a function of the red shift z has to be applied. In summary, the apparent magnitude m is expressed as

$$\mu = m - M = 5 \log_{10} \left(\frac{d_L}{\text{Mpc}} \right) + 25 + K(z). \tag{9.8}$$

The quantity μ is referred to as the *distance modulus*.

We assume that the deviation from uniformity is so small that the density fluctuations can be treated perturbatively. Specifically, we write the density function as

$$\rho(t, \mathbf{r}) = \rho_0(t) + \delta\rho(t, \mathbf{r}) \tag{9.9}$$

and treat $\delta\rho$ as a small perturbation to ρ_0. This means that the energy–momentum tensor has a small perturbative part that depends on the space coordinate. Because of the Einstein equation Eq. (8.1), it induces a perturbation to the metric tensor. The metric tensor is decomposed into scalar, vector, and tensor components (see Section E.4). The tensor component represents gravitational waves. In the context of this book, only the scalar potential is important. The Robertson–Walker metric is modified in the Newtonian gauge to¹⁾

$$ds^2 = \Big(1 + 2\Psi(t, \mathbf{r})\Big) dt^2 - R^2(t) \Big(1 + 2\Phi(t, \mathbf{r})\Big) \gamma_{ij} dx^i dx^j \tag{9.10}$$

where γ_{ij} are the spatial part of the Robertson–Walker metric. For the Euclidean space in the Cartesian coordinate system, $\gamma_{ij} = \delta_{ij}$. Once we introduce the space-dependent energy and momentum, the Hubble constant, in general, becomes a function of the space coordinates, too. However, in the chosen gauge, the Hubble constant is still a function of time only. Ψ is the Newtonian gravitational potential, and Φ is referred to as the *curvature perturbation*. We assume $\Phi = -\Psi$ for a perfect fluid. Therefore, in the following, the local gravitational potential $\Psi(t, \mathbf{r})$ is the only addition to the metric tensor. As a full relativistic treatment of the fluid dynamics is beyond the scope of this book, we will confine ourselves to nonrelativistic

---

1) The metric is, by definition, invariant under general coordinate transformations. But not all of them are physical. For instance, physics law does not change against translaltional shift. One has freedom to change this part freely, which is referred to as the *gauge transformation*. It is often convenient to adopt a certain gauge to clarify physical meaning of the object we are dealing with. See also the Appendix E.3.

treatments and comment on the relativistic extension only when appropriate. In order to understand how the structure formation has proceeded, we will start our discussion with the phenomenological treatment of a nonuniform fluid and how it evolves in the expanding universe.

We first quantify fluctuation of the matter density distribution away from its average over the entire universe. We define the scale-dependent density fluctuation as follows: We measure the average density in a volume with radius $r = \lambda/2$ and define

$$\left.\frac{\delta\rho}{\rho}\right|_{r=\lambda/2} = \frac{\sqrt{<(\rho-\rho_0)^2>}}{\rho_0} \tag{9.11}$$

where the bracket means the statistical average over a large sample of volumes arbitrarily picked up from the space. For $\lambda \gg 200$ Mpc, the fluctuation is small ($\delta\rho/\rho \ll 1$). The fluctuation becomes larger as the scale becomes smaller. At the scale $\lambda \simeq 20$ Mpc, which is the average size of clusters of galaxies, the fluctuation is close to unity, $\delta\rho/\rho{\sim}1$; that is, a mass concentration twice that of the cosmic average is observed (see Eq. (9.66)). At the smaller scale, the universe is even more clumpy by the formation of stars and galaxies, and the density fluctuation is in the nonlinear regime, meaning that the first-order perturbation does not work. The matter is strongly coupled, and a simple analytical treatment is difficult. For instance, a typical galaxy size is ~30 kpc, and has mass $M_G{\sim}2\times10^{11}M_\odot$. Let us find the scale λ_G having the average cosmic density but with the same weight as the galaxy. It is given by

$$\begin{aligned} M_G &= \frac{4\pi}{3}\left(\frac{\lambda_G}{2}\right)^3 \rho_{m0} \\ \rightarrow \quad \lambda_G &\simeq 1.11\,\text{Mpc}\,(\Omega_{m0}h^2)^{-\frac{1}{3}} = 2.2\,\text{Mpc}\left(\frac{\Omega_{m0}h^2}{0.133}\right)^{-\frac{1}{3}} \end{aligned} \tag{9.12}$$

As the size of a typical galaxy is ~30 kpc, one sees that at this scale $\delta\rho/\rho \gg 1$. Once the density fluctuation enters the nonlinear regime ($\delta\rho/\rho > 1$), the nonlinearity grows rapidly. In the standard cosmology, the seed of the fluctuation is assumed to be the quantum fluctuation of the inflaton field. They are assumed to be stretched kinematically by the cosmic expansion and, after the matter–radiation equality epoch, begin to grow dynamically by the gravitational force and acquire the present form. Notice, however, that the earliest fluctuation we can observe is that of the CMBR (cosmic microwave background radiation) or simply CMB. Our discussions will concentrate on the dynamical growth of the fluctuations.

In the following, we use the word "scale" in two ways. One is the scale just mentioned. It is denoted as $r = \lambda/2$ (or sometimes $r = 1/k = \lambda/2\pi$ (which differs by π) is used) and represents the size of objects in question expressed in comoving coordinates. Notice that the comoving variables are independent of the cosmic expansion. The other is the cosmic scale $R(t)$, which grows with time like $\sim t^{2/3}$ in the matter-dominated universe. It is customary to adopt a convention to set

$R_0 = R(t_0) = 1.$[2] In this case, the scale in the comoving coordinate system is quantified by using the same measure as we use at the present time. For instance, the horizon at the time of recombination, expressed in the comoving coordinates, is 1100 times larger than it was at the time of recombination. However, we will retain the explicit notation R_0 as much as possible to clarify what we are referring to.

9.1.2
Boltzmann Equation

Seeds of fluctuations as small as $\sim 10^{-5}$ were found in the CMB by the COBE (cosmic microwave background explorer) in 1992 [751]. It means that the deviation from uniformity at the time of the photon decoupling is so tiny that the perturbational approach to the growth of the fluctuation is justified. We start considering flows of matter taken as a perfect fluid (see Eq. (8.21)). As the Friedmann equation is essentially Newtonian for scales much smaller than the Hubble distance H^{-1} (the size of the horizon), we may start from the Newtonian treatment of the fluid flow. Denoting the density field, velocity field, pressure field and gravitational potential field as $\rho(t, \mathbf{r})$, $\mathbf{v}(t, \mathbf{r})$, $P(t, \mathbf{r})$, $\phi(t, \mathbf{r})$ respectively, the relevant equations are

1) The continuity equation:

$$\dot{\rho} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{9.13}$$

where over-dot in $\dot{\rho}$ means time derivative.

2) The Euler equation of hydrodynamics:

$$\dot{\mathbf{v}} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\left(\nabla\phi + \frac{\nabla P}{\rho}\right). \tag{9.14}$$

3) The Poisson equation of Newtonian gravity:

$$\nabla^2 \phi = 4\pi G_N \rho. \tag{9.15}$$

The above equations hold for a stationary universe. In the expanding universe, it is convenient to work in the comoving coordinate $\boldsymbol{x}$. It is related to the stationary coordinate $\mathbf{r}$ by $\boldsymbol{x} = \boldsymbol{r}/R(t)$. As $\dot{\mathbf{r}} = d(R\mathbf{x})/dt = H\mathbf{r} + R\dot{\mathbf{x}}$, the time variation of a position in the stationary frame consists of an apparent receding motion due to the Hubble expansion and a motion of its own in the comoving coordinate frame multiplied by a scale factor. The latter is referred to as the *peculiar velocity*. Therefore, the velocity field in the stationary frame should be converted to

$$\mathbf{v} \to \mathbf{v} + \dot{R}\mathbf{x}. \tag{9.16}$$

Functions expressed in terms of the stationary coordinate $f(t, \mathbf{r})$ are modified to $f(t, R\mathbf{x}) = F(t, \boldsymbol{x})$. Derivatives in the stationary frame are transformed to

$$\frac{\partial}{\partial t} \to \frac{\partial}{\partial t} - \frac{\dot{R}}{R}\boldsymbol{x} \cdot \nabla_x = \frac{\partial}{\partial t} - H(\boldsymbol{x} \cdot \nabla_x), \qquad \nabla_r \to \frac{1}{R(t)}\nabla_x. \tag{9.17}$$

---

[2] In this case, R_0 is absorbed in the definition of r so that the curvature has to be modified from $k \to K = k/R_0^2$. See footnote 4) in Chapter 8.

In the comoving coordinate system, Eq. (9.13)–(9.15) are expressed as

$$\dot{\rho} + 3H\rho + \frac{1}{R}\nabla \cdot (\rho\, \boldsymbol{v}) = 0 \tag{9.18a}$$

$$\dot{\boldsymbol{v}} + H\boldsymbol{v} + \frac{1}{R}(\boldsymbol{v} \cdot \nabla)\, \boldsymbol{v} = -\frac{1}{R}\nabla\Psi - \frac{\nabla P}{R\rho}, \tag{9.18b}$$

$$\Psi = \phi + \frac{1}{2}R\ddot{R}|\boldsymbol{x}|^2 \tag{9.18c}$$

$$\nabla^2\phi = 4\pi G_N R^2 \rho \tag{9.18d}$$

where we have dropped the suffix x to denote the derivative in the comoving coordinate. Notice that the gravitational potential in the comoving coordinate system is different from that in the stationary frame. The reason why we have modified the gravitational potential as given by Eq. (9.18c) is that the comoving coordinate system and the stationary coordinate system are relatively accelerating. At the observation point, there is no acceleration, but the Hubble expansion generates it at a finite distance. The second term in Eq. (9.18c) represents the inertial force due to the Hubble expansion.

For the uniform universe, $\ddot{R}$ is given by Eq. (8.16)

$$\frac{\ddot{R}}{R} = -\frac{4\pi G_N}{3}\rho_0.^{3)} \tag{9.19}$$

Using Eq. (9.19) and (9.18d), one obtains

$$\nabla^2\Psi = 4\pi G_N R^2(\rho - \rho_0) = 4\pi G_N R^2 \rho_0 \delta \tag{9.22}$$

where δ is defined by $\delta = \delta\rho/\rho$. That is, a uniform energy distribution in the comoving coordinate system does not contribute to the gravitational potential. The potential is sensitive only to fluctuations of the density distribution.

We write the first-order perturbation as

$$\rho(t, \boldsymbol{x}) = \rho_0(t) + \delta\rho(t, \boldsymbol{x}) \equiv \rho_0(t)[1 + \delta(t, \boldsymbol{x})], \tag{9.23a}$$

$$P(t, \boldsymbol{x}) = P(t) + \delta P(t, \boldsymbol{x}) = P(t) + c_s^2 \rho_0(t)\delta(t, \boldsymbol{x}). \tag{9.23b}$$

The second equality for the pressure follows from our assumption of a perfect fluid in which the motion is adiabatic; that is, the entropy along the motion is

---

3) The pressure term was neglected because the Newtonian limit is considered here. For the radiation-dominated universe, the pressure term cannot be neglected. In that case,

$$\frac{\ddot{R}}{R} = -\frac{4\pi G_N}{3}(\rho_0 + 3P_0). \tag{9.20}$$

The Poisson equation, too, has to be modified to

$$\nabla^2\phi = 4\pi G_N R^2(\rho + 3P) \tag{9.21}$$

which is the correct Newtonian limit of general relativity.

conserved. Then the equation of state for the pressure is a function of the density alone and

$$c_s^2 = \left(\frac{\partial P}{\partial \rho}\right)_{\text{adiabatic}} \tag{9.24}$$

is the square of the sound velocity.

As the velocity in the comoving coordinate system appears as a deviation from the uniform universe, it is also a small number. Therefore, we can consider that $|\delta|$, $\boldsymbol{v}$, and $|\Psi|$ are small numbers. Substituting Eq. (9.23a) and (9.23b) into Eq. (9.18a) and (9.18b), the first-order perturbation equations become

$$\dot{\delta} + \frac{1}{R}\nabla \cdot \boldsymbol{v} = 0, \tag{9.25a}$$

$$\dot{\boldsymbol{v}} + H\boldsymbol{v} = -\frac{1}{R}\left(\nabla\Psi + c_s^2\nabla\delta\right). \tag{9.25b}$$

As the transverse (or rotational) component of the vectors are decoupled from the longitudinal component and is a decreasing function of time (see boxed paragraph), we can assume that the velocity does not have transverse components.

Proof that rotational component of the velocity can be dropped. Using $H = \dot{R}/R$ and the fact that ∇ can be replaced with $i\boldsymbol{k}$ in the Fourier space, Eq. (9.25b) can be rewritten as

$$\frac{\partial}{\partial t}(R\boldsymbol{v}) = -i\boldsymbol{k}\left(\Psi + c_s^2\delta\right). \tag{9.26}$$

If we decompose

$$\boldsymbol{v} \equiv \boldsymbol{v}_{\parallel} + \boldsymbol{v}_{\perp} = \frac{\boldsymbol{k}}{|\boldsymbol{k}|^2}(\boldsymbol{k}\cdot\boldsymbol{v}) + \left[\boldsymbol{v} - \frac{\boldsymbol{k}}{|\boldsymbol{k}|^2}(\boldsymbol{k}\cdot\boldsymbol{v})\right]$$

$$\boldsymbol{k}\cdot\boldsymbol{v}_{\perp} = 0, \qquad \boldsymbol{k}\times\boldsymbol{v}_{\parallel} = 0 \tag{9.27}$$

Eq. (9.26) can be decomposed into two independent equations.

$$\frac{\partial(R\boldsymbol{v}_{\perp})}{\partial t} = 0, \qquad \rightarrow \boldsymbol{v}_{\perp} \propto \frac{1}{R(t)}, \tag{9.28a}$$

$$\frac{\partial(R\boldsymbol{v}_{\parallel})}{\partial t} = -i\boldsymbol{k}\left(\Psi + c_s^2\delta\right). \tag{9.28b}$$

We see that the rotational (transverse) component of the velocity is decoupled from the density fluctuations and decreases as $1/R(t)$, which justifies our dropping it from the equation. It also means that the velocity that couples to the density fluctuation can be expressed using a scalar function ψ such that $\boldsymbol{v}\sim\nabla\psi$. It is convenient and conventional to define ψ by

$$\boldsymbol{v} \equiv -\frac{\nabla\psi}{RH}. \tag{9.29}$$

By differentiating Eq. (9.25a) and using Eqs. (9.22) and (9.25b), we obtain the equation for δ in Fourier space.

$$\ddot{\delta}(t,\boldsymbol{k}) + 2H(t)\dot{\delta}(t,\boldsymbol{k}) + \left(\frac{k^2 c_s^2}{R^2(t)} - 4\pi G_N \rho_0\right)\delta(t,\boldsymbol{k}) = 0 \tag{9.30}$$

The sound velocity comes in because the fluctuation is a combined act of compression by gravity and expansion by the pressure. For a relativistic fluid ($P = \rho/3$), the sound velocity is given by

$$c_s = \sqrt{\frac{1}{3}} \qquad \text{: relativistic fluid} \tag{9.31}$$

while for a nonrelativistic fluid, that is, hydrogen atom gas, we have

$$c_s = \sqrt{\gamma\frac{P}{\rho}} = \sqrt{\frac{5T}{3m_N}} \tag{9.32}$$

where m_N is the nucleon mass and we used $PV = N_N T$, $\rho V = N_N m_N$ in deriving the last equality. N_N is the number of nucleons in the volume V. $\gamma = C_P/C_V$ is the adiabatic index, that is, the ratio of the heat capacity at constant pressure to the heat capacity at constant volume. $\gamma = 5/3$ holds for an ideal gas.

We see that the sign of the third term in Eq. (9.30)

$$\kappa_J \equiv \frac{k^2 c_s^2}{R^2(t)} - 4\pi G_N \rho_0 \tag{9.33}$$

determines whether the solution is a growing or damped oscillation. The wave vector $\boldsymbol{k}$ corresponds to the wavelength in the comoving coordinate system or to the physical wavelength by

$$\lambda = \frac{2\pi}{|\boldsymbol{k}|} = \frac{\lambda_{\text{physical}}}{R(t)}. \tag{9.34}$$

One sees that only fluctuations with scales larger than the Jeans length defined by

$$\lambda_J \equiv c_s\sqrt{\frac{\pi}{G_N\rho_0}} \quad \text{or} \quad \lambda_J' \equiv \frac{\lambda_J}{2\pi} = \sqrt{\frac{2}{3}}\frac{c_s}{H} \simeq \frac{c_s}{H} \tag{9.35}$$

can grow. The above expression shows that the Jeans length is the distance that the sound can travel within the Hubble time. We sometimes refer to the *sound horizon*, which is defined by

$$r_s \equiv R(t)\int_0^t \frac{c_s dt}{R(t)} = R(t)\int_0^R \frac{c_s(t)dR'}{R'^2 H(R')} \tag{9.36}$$

which is the horizon with the light velocity replaced with the sound velocity.

Jeans mass: Here we describe the physical picture of the Jeans mass more intuitively. In order for the mass fluctuation to grow, it has to overcome the pressure to disperse it. There exists a minimum amount of mass (Jeans mass M_J) to make the attractive gravitational force to overcome the pressure. Stated in different words, the in-fall velocity due to the gravitational force has to be larger than the sound velocity c_s. Writing the condition for the kinetic energy of particles inside a sphere of diameter λ'_J to lose against the gravitational potential energy,

$$\frac{c_s^2}{2} < \frac{G_N M_J}{\lambda'/2} = \frac{G_N}{\lambda'/2}\frac{4\pi}{3}\left(\frac{\lambda'}{2}\right)^3 \rho = \frac{\pi G_N \rho \lambda_J'^{\,2}}{3}. \tag{9.37}$$

Using Eq. (8.10), we have

$$\lambda'_J = c_s \left(\frac{2\pi G_N \rho}{3}\right)^{-\frac{1}{2}} = \frac{2c_s}{H} \tag{9.38}$$

which approximately reproduces Eq. (9.35).

One can also define the Jeans mass of the baryon by

$$M_J = \frac{4\pi}{3}\rho_b\left(\frac{\lambda_J}{2}\right)^3 = \frac{8}{9}\sqrt{\frac{2}{3}}\pi^4\rho_b\left(\frac{c_s}{H}\right)^3. \tag{9.39}$$

To convert the Jeans wavelength into the comoving coordinate system, one has to multiply it by the cosmic scale $R(t)$, but it is compensated by $\rho_b \sim R^{-3}$. Therefore, the Jeans mass is independent of the cosmic expansion.

In the radiation-dominated era, the sound velocity is $c_s = 1/\sqrt{3}$. Therefore, the Jeans wavelength is larger than the horizon at the time.

$$\lambda_J = 2\pi\sqrt{\frac{2}{3}}\frac{c_s}{H} = \frac{4\sqrt{2}\pi}{3}t > d_H(t) \tag{9.40}$$

where we used $H^{-1} = 2t$ for radiation-dominated era. Strictly speaking, the Newton's law cannot be applied for such a large scale, but the conclusion is the same. That is, in the radiation-dominated universe, the density fluctuation cannot grow. The rule applies to any matter particles. For the baryons, the rule is even more severe. As they are strongly coupled with photons via electrons, the baryonic density fluctuation, if generated, cannot grow because of the photon's pressure until the photon is decoupled from the baryons. Until then, it can only propagate as a sound wave.

After the decoupling, the pressure is provided by the baryon's own kinetic energy. Using Eq. (9.32), the sound velocity becomes $c_s^2 = (5/3)T_B/m_N$, where T_B denotes the baryon temperature. The baryon temperature behaves $T_B \sim \langle p^2 \rangle \sim R^{-2} \sim T^2$

and can be expressed as

$$T_B = T^2/T_{\mathrm{dc}} \quad \rightarrow \quad c_s^2 = \frac{5}{3}\frac{T^2}{m_N T_{\mathrm{dc}}}. \tag{9.41}$$

Therefore, the Jeans mass of the baryon can be calculated as

$$M_J = \frac{4}{3}\pi\rho_b\left(\frac{\lambda_J}{2}\right)^3 \overset{(9.35)}{=\!=} \frac{\pi^{\frac{5}{2}}}{6}\rho_b\left(\frac{c_s^2}{G_N\rho_0}\right)^{\frac{3}{2}}. \tag{9.42a}$$

For simplicity, we approximate $\rho_0 \simeq \rho_b$ and express the Jeans mass as

$$M_J = \frac{5}{18}\sqrt{\frac{5}{3}}\pi^{\frac{5}{2}}\left(\frac{1}{G_N m_N}\right)^{\frac{3}{2}}\left(\frac{T^3}{\rho_b}\right)^{\frac{1}{2}}\left(\frac{T}{T_{\mathrm{dc}}}\right)^{\frac{3}{2}}. \tag{9.42b}$$

As $T^3/\rho_b{\sim}1/(\rho_b R^3)$ is a conserved number, it can be replaced with

$$\frac{T^3}{\rho_b} = \frac{T_0^3}{\rho_{b0}} = \frac{T_0^3}{\Omega_{b0}\rho_{c0}} \tag{9.43}$$

where $\rho_{c0} = 1.0537 \times 10^{-5}h^2\mathrm{GeV\,cm^{-3}}$ is the critical density at present time. The Jeans mass of the baryon can be expressed as

$$\begin{aligned} M_J &= \frac{5}{18}\sqrt{\frac{5}{3}}\pi^{\frac{5}{2}}\frac{1}{\sqrt{\Omega_{b0}h^2}}\left(\frac{T_0}{m_N}\right)^{\frac{3}{2}}\frac{M_{\mathrm{Pl}}^3}{(\rho_{c0}/h^2)^{\frac{1}{2}}}\left(\frac{T}{T_{\mathrm{dc}}}\right)^{\frac{3}{2}} \\ &\simeq 1.4 \times 10^5(\Omega_{b0}h^2)^{-\frac{1}{2}}\left(\frac{T}{T_{\mathrm{dc}}}\right)^{\frac{3}{2}} M_\odot. \end{aligned} \tag{9.44}$$

This Jeans mass is about the size of globular clusters and is much smaller than the galactic mass $M_{\mathrm{galaxy}} \simeq 10^{11}M_\odot$. Thus, the baryon mass fluctuation can grow after the photon decoupling.

In summary, we have shown that, for the mass fluctuation to grow, it is necessary that the baryons be decoupled from the photons. We next investigate if the mass fluctuations have enough time to grow to the present size of the fluctuation.

9.1.3
Growth of the Fluctuation

For small k (or large $\lambda \gg \lambda_J$), terms proportional to k^2 in Eq. (9.30) can be neglected. For flat, matter-dominated universe ($R \propto t^{2/3}$), $H(t) = 2/(3t)$ holds (see Eqs. (8.55) (8.56)). Together with the Friedmann equation, we have $\rho_0 = 1/6\pi G_N t^2$, so that Eq. (9.30) becomes

$$\ddot{\delta}(t,\boldsymbol{k}) + \frac{4}{3t}\dot{\delta}(t,\boldsymbol{k}) - \frac{2}{3t^2}\delta(t,\boldsymbol{k}) = 0 \tag{9.45}$$

which has the solution

$$\delta(t,\boldsymbol{k}) = c_+ t^{2/3} + c_- t^{-1}. \tag{9.46}$$

That is, the growing mode that we are interested in evolves in proportion to the scale R. Notice that, if the universe is stationary, that is, if the second term in Eq. (9.30) is absent and $\dot{R} = \dot{\rho}_0 = 0$, the growth rate will be exponential. The expanding universe ameliorates the growth rate to be linear. The perturbation method works as long as the fluctuation δ is small. If it becomes larger than unity, the equations are no longer linear and numerical N-body simulation has to be carried out. The density fluctuation at a given time is usually assumed to be Gaussian. In fact, it is exactly what the inflation models predict, where the origin of the primordial fluctuation is ascribed to quantum fluctuation of the inflaton field.

The linear growth of the fluctuation $\delta\rho/\rho$ and the uniformity of the CMB arouse difficulty. As already mentioned, the baryon density fluctuation cannot grow before the decoupling. As the present fluctuation is at least as large as $\sim$1, the fluctuation at the time of decoupling is estimated to be at least larger than

$$\delta(t_{\text{dc}}) = \delta(t_0)\frac{R(t_{\text{dc}})}{R(t_0)} \sim \frac{\delta(t_0)}{(1+z_{\text{dc}})} \sim 1 \times 10^{-3}.^{4)} \tag{9.47}$$

On the other hand, the observed temperature fluctuation of the CMB is about $\approx 10^{-5}$. As both entropy and baryon number are conserved, we are considering the adiabatic perturbations and the entropy per baryon $s_b = s/n_b$ is conserved. From $\delta s_b = 0$, one obtains $\delta n_b/n_b = \delta\rho_b/\rho_b = 3\delta T/T$. Then the temperature fluctuation translates to the density fluctuation at most by

$$\delta_{\text{dc}} = \left.\frac{\delta\rho}{\rho}\right|_{\text{dc}} \sim 3\left.\frac{\delta T}{T}\right|_{\text{dc}} \sim 10^{-4}. \tag{9.48}$$

This is in contradiction to Eq. (9.47) because the baryon fluctuation should be equal to the photon energy density fluctuation at the time of decoupling.[5] The solution was provided by the dark matter.

9.1.4
Dark Matter

A detailed discussion of the dark matter will be presented in the next chapter. Here, we only list some of its salient features relevant to the structure formation.

We argued that the baryon fluctuation at the time of decoupling is too small to explain the observed density fluctuation. If there is a matter that does not interact with photons or baryons, its fluctuation can begin to grow at the time of matter–radiation equality. As $R(t_{\text{eq}})/R(t_0) = (1+z_{\text{eq}})^{-1} \sim 3 \times 10^{-4}$, the dark matter can provide the fluctuation of about the right size. When the baryons are decoupled, they are captured by the gravitational field created by the dark matter and their fluctuation catches up fast to become as large as that of the dark matter.

---

4) The existence of the cosmological constant affects the expansion speed of the universe but its effect is small and does not change the conclusion under assumption of the matter dominance.

5) Actually, the photon–baryon fluid fluctuations that existed before the decoupling is almost completely erased by the Silk damping (see Section 9.2.5). Therefore, one may consider that the baryon fluctuation did not exist at the time of decoupling.

This is one reason to require the presence of dark matter. Another theoretical indication comes from the galaxy formation theory, which requires the existence of spherical mass distribution interacting only gravitationally to stabilize the motion of spiral galaxies. Observationally, evidences for the existence of the dark matter are overwhelming. Its existence can be detected directly through gravitational forces or indirectly from M/L, that is, the mass to luminosity ratio of astronomical objects. The large-scale structure of the universe and the observed temperature anisotropy fits well with the ΛCDM model which has become the standard of present cosmology. It takes into account of the dark energy in the cosmic energy composition. Details of the dark energy will be discussed in the next chapter. The dark matter occupies about 25% of the total cosmic energy budget. The baryon is known to occupy only ~5%. It follows that the dark matter is not an ordinary particle known in the Standard Model (SM). It must be a stable particle and neutral colorwise as well as electrically. The best candidate is the LSP (Lightest Supersymmetric Particle), LKP (Lightest KK Particle) or the axion.

9.1.5
Jeans Wavelength of the Neutrino

One candidate of the dark matter in the past was the neutrino. We will calculate the Jeans mass of the neutrino and see whether it fits with the dark matter.

The neutrino interaction is weak, almost collision-less with other particles. Therefore, it is freely streaming in the universe. In this case, the sound velocity is replaced with the dispersion velocity. This means that the Jeans length of the neutrino $\lambda_{J\nu}$ can be defined as the maximum distance that the neutrino can travel before $t \leq t_{\text{eq}}$. Once the universe becomes matter-dominated, the density fluctuation begins to grow. The neutrino temperature at the time of matter–radiation equality is slightly lower than that of the photon and is given by (see 8.120)

$$\begin{aligned} T_{\nu\,\text{eq}} &= T_{\text{eq}}\frac{T_\nu}{T_\gamma} = T_0(1+z_{\text{eq}})\frac{T_\nu}{T_\gamma} \overset{(8.134)}{=\!=} 2.348\times 10^{-4}\ \text{eV}\ \times 3200\times 0.71 \\ &= 0.53\ \text{eV} \end{aligned} \tag{9.49}$$

which is comparable with the current upper limit of the neutrino mass. Therefore, we may consider that the neutrino remains relativistic at $t = t_{\text{eq}}$. Assuming the neutrino mass to be much smaller than 1 eV, it becomes nonrelativistic after t_{eq}. As the relativistic neutrino diffuses with the light velocity, the diffusion length is given by the horizon. Therefore, we may approximate the Jeans length by the horizon at the time when the neutrino becomes nonrelativistic assuming matter dominance. In the comoving coordinate system, it is given by

$$\lambda_{J\nu} \simeq (1+z_{\rm NR})d_H(z_{\rm NR}) \overset{(8.77)}{=\!=} \frac{R_0}{R(t_{\rm NR})}R(t_{\rm NR})\int_0^{t_{\rm NR}}\frac{dt'}{R(t')} = \int_{z_{\rm NR}}^{\infty}\frac{d\tilde{z}}{H(\tilde{z})}$$
$$\overset{(8.68)}{\simeq} \frac{1}{H_0\sqrt{\Omega_{m0}}}\int_{z_{\rm NR}}^{\infty}\frac{d\tilde{z}}{(1+\tilde{z})^{\frac{3}{2}}} = \frac{2}{H_0\sqrt{\Omega_{m0}}}(1+z_{\rm NR})^{-\frac{1}{2}}. \tag{9.50}$$

If we set $T_{\nu,\rm NR} = m_\nu/3$ as the temperature when the neutrino becomes nonrelativistic, then

$$1+z_{\rm NR} = \frac{T_{\rm NR}}{T_0} = \frac{T_{\nu,\rm NR}}{T_0}\frac{T_\gamma}{T_\nu} = \frac{m_\nu}{T_0}\frac{1}{3}\left(\frac{11}{4}\right)^{\frac{1}{3}}. \tag{9.51}$$

Putting the number in Eqs. (9.50) and (9.51), we obtain

$$\lambda_{J\nu} = 134\left(\frac{\sum_i m_{\nu i}}{\rm eV}\right)^{-\frac{1}{2}}\frac{1}{\sqrt{\Omega_{m0}h^2}}\,\rm Mpc$$
$$\text{or}\quad k_{\nu\rm NR} = 0.024\left(\frac{\sum_i m_{\nu i}}{1\,\rm eV}\right)^{\frac{1}{2}}\left(\frac{\Omega_{m0}}{0.27}\right)^{-\frac{1}{2}}\ (\rm h\ Mpc^{-1}) \tag{9.52}$$

where $\sum_i$ takes into account the three flavors of the neutrino in the SM.

The scale is typically larger than a cluster's size. The dark matter that is relativistic at time $t = t_{\rm eq}$ is referred to as the *hot dark matter* (HDM). In the HDM scenario, the structure formation begins at the scale of clusters. Later, these structures are considered to fragment and generate galaxies. This is a top-down scenario as contrasted to the CDM (cold dark matter) scenario which is a bottom-up process. In the HDM scenario, we may have structures at the level of clusters or superclusters but there is not enough time for them to decompose and to form galaxies. Besides, the observed statistical properties are different from those derived using the HDM hypothesis. Hence it is believed that our universe is not dominated by the HDM. In the CDM scenario, the dark matter is already nonrelativistic at $t = t_{\rm eq}$. It is possible to start clustering from as small as the globular cluster, as shown by the baryon's Jeans mass (Eq. (9.44)).

A representative of the HDM is the neutrino, and those of CDM include almost all other particles. Among them, the most popular candidates are the LSP, the LKP, and the axion. The axion has a very small mass, but behaves as a Bose-condensed fluid with vanishing momentum and is a member of CDM (see Section 7.4.2).

9.1.6
Power Spectrum

Here we try to understand the observed large-scale structure of the universe, that is, distribution of galaxies shown in Figure 9.1, and investigate the power spectrum of the density fluctuations. Expressing the density as a function of space points $\boldsymbol{r}$, its fluctuation is defined by

$$\begin{aligned}\delta(\boldsymbol{r}) &\equiv \frac{\delta\rho}{\rho} = \frac{\rho(\boldsymbol{r})-\rho_0}{\rho_0}\\ &= \sum_{\boldsymbol{k}} \delta(\boldsymbol{k}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \xrightarrow{V\to\infty} \frac{V}{(2\pi)^3}\int d^3k\delta(\boldsymbol{k})e^{i\boldsymbol{k}\cdot\boldsymbol{r}}.\end{aligned} \tag{9.53}$$

The variable that is directly measured is the mean-square average of the fluctuation.

$$\begin{aligned}<\delta(\boldsymbol{r})^2> &= \frac{1}{V}\int dV\,|\delta(\boldsymbol{r})|^2 = \frac{V}{(2\pi)^3}\int d^3k\,|\delta(\boldsymbol{k})|^2\\ &= V\int_0^\infty \frac{k^3|\delta_k|^2}{2\pi^2}\frac{dk}{k} \equiv \int_0^\infty \Delta_k^2 \frac{dk}{k}\\ \Delta_k^2 &= V\frac{k^3}{2\pi^2}|\delta_k|^2\end{aligned} \tag{9.54}$$

where for isotropic distributions we have written $\delta(\boldsymbol{k}) = \delta(|\boldsymbol{k}|) = \delta_k$. $P(k) \equiv \delta_k^2$ is referred to as the *power spectrum* and represents the density fluctuation at scale $r_0 = 1/k$ in real space.

Let us see how δ_k is related to the fluctuation in real space. Suppose we sample a total of N spherical volumes with radius $r_0 = 1/k = \lambda/2\pi$ and measure the total mass inside the jth sphere to be M_j. We define a window function by

$$W(\boldsymbol{r}) = \begin{cases} 1 & r \le r_0 \\ 0 & r > r_0 \end{cases}, \qquad V_W = \int d\boldsymbol{r}\,W(\boldsymbol{r}) = \frac{4\pi}{3}r_0^3, \tag{9.55a}$$

$$M(\mathbf{x}_j) \equiv M_j = \int d\boldsymbol{r}\,\rho(\mathbf{x}_j + \boldsymbol{r})W(\boldsymbol{r}), \tag{9.55b}$$

$$\overline{M} =< M(\mathbf{x}_j) >= \frac{1}{N}\sum_{j=1}^{N} M_j. \tag{9.55c}$$

We define the mass fluctuation at scale r_0 by

$$\begin{aligned}\delta_M(\mathbf{x}_j) &= \frac{1}{V_W}\int d\boldsymbol{r}\;W(\boldsymbol{r})\frac{\rho(\boldsymbol{r}+\mathbf{x}_j)-\rho_0}{\rho_0} = \frac{1}{V_W}\int d\boldsymbol{r}W(\boldsymbol{r})\delta(\mathbf{x}_j+\boldsymbol{r})\\ \therefore\quad \delta_M^2 &= \left\langle\left(\frac{1}{V_W}\int d\boldsymbol{r}\delta(\mathbf{x}_j+\boldsymbol{r})W(\boldsymbol{r})\right)^2\right\rangle.\end{aligned} \tag{9.56}$$

Substituting the Fourier transform of the window function

$$W(\boldsymbol{r}) = \frac{V_W}{(2\pi)^3}\int d\boldsymbol{k}\;W(\boldsymbol{k})e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \tag{9.57}$$

into Eq. (9.56), we have

$$\delta_M(\mathbf{x}_j) = \int d\boldsymbol{r}\;\delta(\mathbf{x}_j+\boldsymbol{r})W(\boldsymbol{r}) = \frac{V_W}{(2\pi)^3}\int d\boldsymbol{k}\;\delta_k W_k e^{i\boldsymbol{k}\cdot\mathbf{x}_j} \tag{9.58}$$

where $W_k = W(|\boldsymbol{k}|) = W(\boldsymbol{k})$ for isotropic functions. If we integrate over $\mathbf{x}_j$ to make the average, δ_M is the same as $\delta(\boldsymbol{r})$, with δ_k simply replaced by $\delta_k W_k$. In reality, we cannot integrate over the whole space to make the average but take N samples and make their statistical average as given by Eq. (9.55c). If we can take a large number of independent statistical samples, we are justified in approximating the statistical average by the space integral. Then, using Eq. (9.54) we have

$$\delta_M^2 = \int_0^\infty \Delta_k^2 W_k^2 \frac{dk}{k}. \tag{9.59}$$

For the distribution Eq. (9.55), the Fourier transform of the window function becomes

$$W_k = \frac{1}{V_W} \int d\boldsymbol{r}\, W(\boldsymbol{r}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = 3\left[\frac{\sin kr_0 - kr_0 \cos kr_0}{(kr_0)^3}\right] = 3\frac{j_1(kr_0)}{kr_0} \tag{9.60}$$

where $j_1(x)$ is the spherical Bessel function of the first order. Then

$$\delta_M^2 = \int \frac{dk}{k} \Delta_k^2 W_k^2 = \int \frac{dk}{k} \Delta_k^2 \left(\frac{3j_1(kr_0)}{kr_0}\right)^2. \tag{9.61}$$

A model often used in expressing the power spectrum assumes

$$|\delta_k|^2 \sim k^n \quad \text{or} \quad \Delta_k^2 = Ak^{n+3}. \tag{9.62}$$

Furthermore, one often uses the approximation

$$3\frac{j_1(z)}{z} = \begin{cases} 1 & z \le 1 \\ 0 & z > 1. \end{cases} \tag{9.63}$$

Then, for $\Delta_k^2 = Ak^{n+3}/2\pi^2$, one has

$$\delta_M^2 = \frac{\Delta^2(k = 1/r_0)}{n+3} \sim Ak^{n+3}\Big|_{k=1/r_0}. \tag{9.64}$$

That is, if $n > -3$, the mass fluctuation at scale $r_0 = 1/k$ is directly proportional to Δ_k^2. As $W_k^2 \propto k^{-4}$ for $k \to \infty$, we need $n < 1$ in order to be able to neglect the short wavelength components. But conventionally, one does not worry about it, because ultraviolet convergence is satisfied by choosing the window function appropriately, for instance, a Gaussian distribution $W \sim \exp(-r^2/2r_0^2)$.

Equation (9.64) means that the mass fluctuation δ_M evaluated at scale $r_0 = 1/k$ is directly proportional to the power spectrum (times k^3)[6], that is, we can express the magnitude of the $\delta(\boldsymbol{r})$ at scale $1/k$ by

$$|\delta(\boldsymbol{r})|_{r_0=1/k} \simeq |\delta_M|_{r_0} \sim \Delta_k. \tag{9.65}$$

According to observations, $\delta_M = 1$ is reached approximately at the scale $r_0 = 8\text{h}^{-1}$ Mpc (see Eq. (9.66)). The density fluctuation at $r_0 = 8\text{h}^{-1}$ Mpc is given a special name

$$\sigma_8 \equiv \delta_M|_{r_0=8h^{-1}\text{Mpc}} = 0.81 \pm 0.03 \quad [7]. \tag{9.66}$$

It is conventionally adopted as the value of the density fluctuation at which the structure formation enters the nonlinear regime. Its value is traditionally used to determine the overall normalization for the power spectrum.

---

[6] This is why some people define the power spectrum not by δ_k^2 but by Δ_k^2.

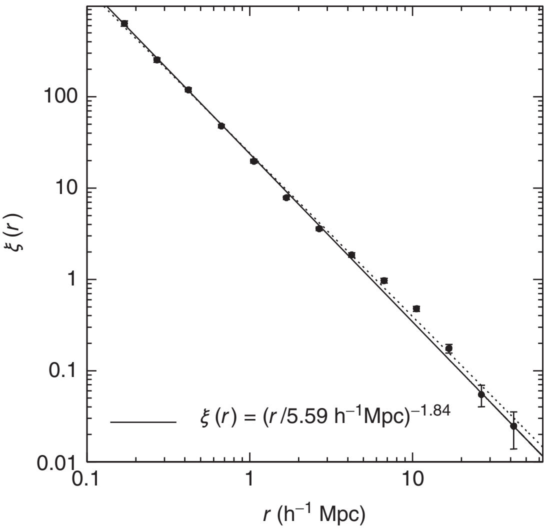


Figure 9.2 Real space correlation function $\xi(r)$ for the flux-limited galaxy sample. The solid and dotted lines show the corresponding power-law fits. (Reproduced with permission of [752].)

Galaxy correlation We show that the density fluctuation is directly related to the correlation of galaxy distributions. The correlation function $\xi(\boldsymbol{x})$ is defined by

$$\begin{aligned} &\xi(\boldsymbol{x}) =<\delta(\boldsymbol{x}+\boldsymbol{r})\delta(\boldsymbol{r})>= \frac{V}{(2\pi)^3}\int d\boldsymbol{k}\,|\delta_k|^2 e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \\ \therefore\quad &|\delta_k|^2 = \frac{1}{V}\int d\boldsymbol{r}\,\xi(\boldsymbol{r})e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \end{aligned} \tag{9.67}$$

With a plausible assumption that the mass density of the galaxy is proportional to its number density of luminous galaxies,[7] the number of galaxies in a spherical shell (volume $\delta V = 4\pi r^2 dr$) is given by

$$n(\boldsymbol{r}) = \overline{n}\,\delta V\Big(1+\xi(r)\Big) \tag{9.68}$$

where we have assumed isotropic distribution $\xi(\boldsymbol{r}) = \xi(r)$. Observations show (Figure 9.2 [753, 754]) that

$$\xi_{GG}(r) \simeq [(r/5.4\,\mathrm{h}^{-1}\,\mathrm{Mpc})]^{-1.8} \qquad r \lesssim 10\,\mathrm{h}^{-1}\mathrm{Mpc} \tag{9.69a}$$

$$\xi_{CC}(r) \simeq [(r/25\,\mathrm{h}^{-1}\,\mathrm{Mpc})]^{-1.8} \qquad r \lesssim 100\,\mathrm{h}^{-1}\mathrm{Mpc} \tag{9.69b}$$

where ξ_{GG} and ξ_{CC} are the galaxy–galaxy and cluster–cluster correlation functions. It is interesting to note that the cluster correlation exhibits the same power law as that of the galaxy correlation at widely different scales.

---

[7] The assumption is being confirmed by comparing the distributions of luminous galaxies with those of the dark matter obtained by gravitational lensing.

Setting $|\delta_k|^2 = Ak^n$, we have

$$\xi(r) = \frac{V}{(2\pi)^3}\int d\boldsymbol{k}\,|\delta_k|^2 e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \frac{AVr^{-n-3}}{2\pi^2}\Gamma(n+2)\sin\frac{n+2}{2}\pi. \tag{9.70}$$

Therefore, the data means $\xi \sim r^{-1.8} \to n = -1.2$. In principle, to obtain the spatial function $\delta(\boldsymbol{r})$, one needs data at $0 < r < \infty$, but in reality it is difficult and so the function $J_3(r_0)$ defined by the following formula is often used:

$$\begin{aligned} J_3(r_0) &\equiv \frac{1}{4\pi}\int_0^{r_0} d\boldsymbol{r}\,\xi(r) = \frac{1}{4\pi}\int d\boldsymbol{r}\,W(\boldsymbol{r})\xi(r) = \frac{VV_W}{4\pi(2\pi)^3}\int d\boldsymbol{k}\,W_k\delta_k^2 \\ &\simeq \frac{VV_W}{4\pi(2\pi)^3}\int d\boldsymbol{k}\,W_k^2\delta_k^2 = \frac{r_0^3}{3}\int\frac{dk}{k}\Delta_k^2 W_k^2 = \frac{r_0^3}{3}\delta_M^2 \end{aligned} \tag{9.71}$$

where Eq. (9.63) was used in going to the second line. If one uses the above function, then

$$\delta_M^2 = \frac{3}{r_0^3}J_3(r_0). \tag{9.72}$$

From the above discussions, one sees that the power spectrum can be reconstructed using density fluctuations and galaxy correlation functions. At scales larger than $\lambda > 200$ Mpc, no structure is seen, confirming the standard cosmological assumptions.[8]

BAO: Baryon Acoustic Oscillation We have seen that, prior to the photon decoupling, baryons were tightly coupled photons through electrons and were oscillating as acoustic waves. It is referred to as the *baryon acoustic oscillation*. The fluctuation does not evolve while it is oscillating. Traces of the oscillation are imprinted in the CMB spectrum and also observed in the density fluctuation of the galaxies. Note that the acoustic baryon oscillation is a phenomenon at $z > 1100$ while the BAO is the corresponding oscillation pattern that appears in the structure formation at $z \sim 0.1 - 0.5$. The trace appears at the scale of acoustic horizon at the time of the photon decoupling $d_{H\,\mathrm{BAO}} \simeq 100\,\mathrm{h}^{-1}$ Mpc in the comoving coordinate system (see Figure 9.3(a), Eq. (9.80)).

The importance of the BAO is that it can serve as a standard rule at the cosmic distance at $z \lesssim 1$ as its dynamics is well understood and both the longitudinal and transverse size of the BAO are known accurately. Figure 9.3(b) certifies that the BAO at two different redshifts agree exactly. Figure 9.4 plots the distance determined by the BAO as a function of z. The agreement of the data with predictions based on ΛCDM model with inputs from WMAP (Wilkinson microwave anisotropy probe) is excellent.

Here the distance is defined by $D_V(z) = \left[\left(c/H(z)\right)\left\{(1+z)D_A(z)\right\}^2\right]^{\frac{1}{3}}$ to account for that the galaxy number density is proportinal to the volume element which is proportional to the square of angular size times the line of sight distance (see Eq.(11.62)). In fact, the most accurate and comprehensive set of data to determine cosmological parameters are a combination of the large-scale structure

---

[8] Some recent SDSS survey analyses indicate deviations from uniformity [755].

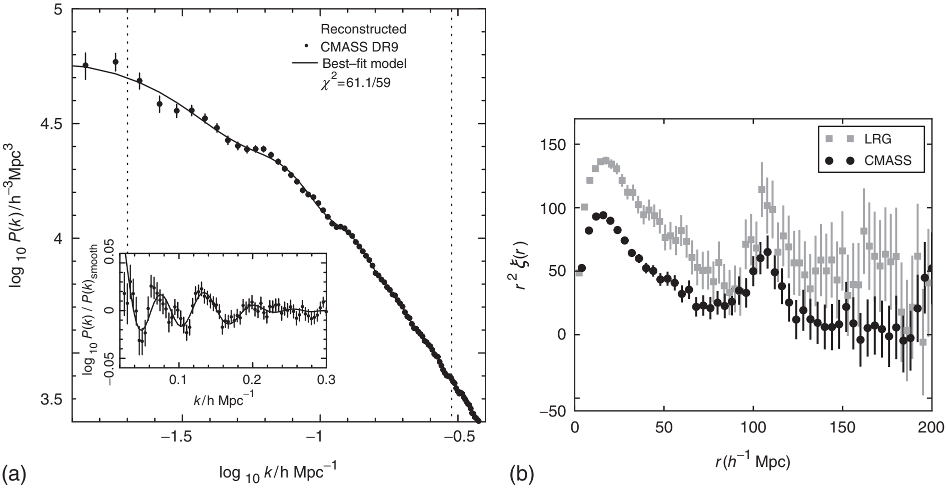


Figure 9.3 (a) Galaxy power spectrum and correlation at $z = 0.57$ from the Baryon Oscillation Spectroscopic Survey (BOSS), which is a part of the Sloan Digital Sky Survey (SDSS III). The best-fit ΛCDM model is shown as a solid line). The BAO inset shows the same data (between the two verticall dotted lines) and model divided by a spline fit to the smooth component. (b) Correlation function measured from the same data (black circles) versus that from SDSS-II Luminous Red Galaxies (LRG) data (gray squares) at $z = 0.35$. The two analyses use slightly different formalism, but one can clearly see that the acoustic peak at two different z are located exactly at the same position. (Reproduced with permission of [756–758].)

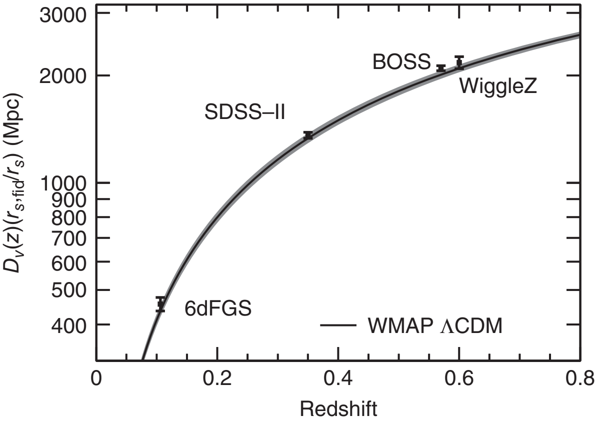


Figure 9.4 A plot of the distance-redshift relation from various BAO measurements from spectroscopic data sets. The grey band is the 1σ prediction from WMAP under the assumption of a flat Universe with a cosmological constant. $(r_{s,fid}/r_s)$ is a correction factor to fit the WMAP prediction. (Reproduced with permission of [756].)

analysis using the BAO and gravitational lensing, the CMB analysis combined with those from SNI (supernovae type Ia observation (see Section 11.1)). It is considered to serve as a useful tool in future in investigating the properties of the dark energy, which will be discussed in detail in Chapter 11.

Sound Horizon at Decoupling Let us calculate the exact position of the BAO. The acoustic oscillation of the baryon-photon fluid begins when its wavelength λ becomes shorter than the horizon size which defines the time zero ($t = 0$) and lasts until the photon decouples. At decoupling, the fluid stops oscillating and its phases is frozen, which is imprinted on the post-decoupling fluctuation as the BAO. The scale of the BAO is determined by the maximum distance that the sound wave can travel by the time of the decoupling, that is, by the acoustic horizon defined by Eq. (9.36). The sound horizon is nothing but the particle horizon with the light velocity replaced with the sound velocity. What we need is the size of the sound horizon at the decoupling $r_s(t_{\rm dc})$ as observed today, that is, the sound horizon in the comoving coordinate system. It is defined by

$$d_s(t_{\rm dc}) \equiv \frac{R_0}{R(t_{\rm dc})} r_s = R_0 \int_0^{t_{\rm dc}} \frac{c_s}{R} dt. \tag{9.73}$$

For a given wavenumber k, the wave behaves as $\sim e^{ikd_s(t)}$. Two waves propagating in opposite directions constitute a stationary wave whose amplitude is expressed as $\sim \cos(kd_s(t) + \alpha)$, where the phase α is determined by the initial condition. For adiabatic fluctuations, the relation $\delta_b = (3/4)\delta_\gamma$ holds, and the fluctuation exists at $t = 0$; hence $\alpha = 0$ (see Eqs. (9.126) and (9.134)). When the photon decouples, the fluctuation with wavenumber k satisfying $kd_s(t_{\rm dc}) = m\pi$ $(m = 1, 2, \cdots)$ is left with the maximum amplitude. That is, we should observe a fluctuation maximum at $d_s(t_{\rm dc}) = m\pi/k$.

The acoustic velocity is given by

$$\begin{aligned} c_s^2 &= \left(\frac{\delta P}{\delta \rho}\right)_s = \left(\frac{\delta(P_b + P_\gamma)}{\delta(\rho_b + \rho_\gamma)}\right)_s = \left(\frac{\delta P_\gamma}{\delta \rho_\gamma}\right)_s \frac{1}{1 + \delta\rho_b/\delta\rho_\gamma} = \frac{1}{3}\frac{1}{1 + R_b} \\ R_b(t) &= \frac{\delta\rho_b}{\delta\rho_\gamma} = \frac{\delta_b}{\delta_\gamma}\frac{\rho_b}{\rho_\gamma} = \frac{3}{4}\frac{\rho_b}{\rho_\gamma}. \end{aligned} \tag{9.74}$$

Substituting Eq. (9.74) in Eq. (9.73), we obtain [759]

$$\begin{aligned} d_s(t) &= \frac{R_0}{\sqrt{3}} \int_0^{R(t)} \frac{1}{\sqrt{1 + R_b}} \frac{dR}{R^2 H} = \frac{\sqrt{2/3}}{k_{\rm eq}\sqrt{R_{b,\rm eq}}} \int_0^{R_b(t)} \frac{dx}{\sqrt{(1 + x)(x + R_{b,\rm eq})}} \\ &= \frac{2\sqrt{2/3}}{k_{\rm eq}\sqrt{R_{b,\rm eq}}} \ln\left[\frac{\sqrt{1 + R_b(t)} + \sqrt{R_b(t) + R_{b,\rm eq}}}{1 + \sqrt{R_{b,\rm eq}}}\right] \end{aligned} \tag{9.75}$$

where $x = R_b(t)$ and we used

$$R_b(t) = \frac{3}{4}\frac{\rho_b}{\rho_\gamma} = \frac{3}{4}\frac{\rho_b}{\rho_m}\frac{\rho_m}{\rho_{\rm rad}}\frac{\rho_{\rm rad}}{\rho_\gamma} = \frac{3}{4}\frac{\rho_b}{\rho_m}\frac{\rho_{\rm rad}}{\rho_\gamma}\frac{\rho_m(t_{\rm eq})}{\rho_{\rm rad}(t_{\rm eq})}\frac{R}{R_{\rm eq}} = R_{b,\rm eq}\frac{R}{R_{\rm eq}} \tag{9.76}$$

and the fact that ρ_b/ρ_m and $\rho_{\text{rad}}/\rho_\gamma$ are both constants.

$$\begin{aligned} R^2H &= \sqrt{\frac{8\pi}{3}G_N(\rho_m+\rho_{\text{rad}})R^4} = \sqrt{\frac{8\pi}{3}G_N\rho_{\text{rad}}R^4\left(1+\frac{\rho_m}{\rho_{\text{rad}}}\right)} \\ &= \frac{H_{\text{eq}}}{\sqrt{2}}R_{\text{eq}}^2\sqrt{\left(1+\frac{R}{R_{\text{eq}}}\right)} = \frac{H_{eq}R_{eq}^2}{\sqrt{2R_{b,eq}}}\sqrt{(x+R_{b,eq})}. \end{aligned} \tag{9.77}$$

Here, $k_{\text{eq}} = R_{\text{eq}}H_{\text{eq}}/R_0$ is the inverse Hubble size at the time of matter–radiation equality defined by Eq. (8.137). Notations are a bit confusing, but note that R_b, $R_{b,\text{eq}}$ denote the baryon-to-photon density ratio to be distinguished from R, R_{eq} which are cosmic scale factors. Using $R_{b,\text{dc}} = R_b(t_{\text{dc}})$, the value of the sound horizon at the photon decoupling is expressed as

$$d_s(t_{\text{dc}}) = \frac{2\sqrt{2/3}}{k_{\text{eq}}\sqrt{R_{b,\text{eq}}}}\ln\left[\frac{\sqrt{1+R_{b,dc}}+\sqrt{R_{b,\text{eq}}+R_{b,dc}}}{1+\sqrt{R_{b,\text{eq}}}}\right]. \tag{9.78}$$

$R_{b,\text{eq}}$ and $R_{b,dc}$ can be determined as follows: Using Eq. (9.76)

$$R_{b,\text{eq}} = R_b(t_{\text{eq}}) = \frac{3}{4}\frac{\rho_b}{\rho_m}\frac{\rho_{\text{rad}}}{\rho_\gamma}\frac{\rho_m(t_{\text{eq}})}{\rho_{\text{rad}}(t_{\text{eq}})} = \frac{3}{4}\frac{\rho_{rad\,0}}{\rho_{\gamma 0}}\frac{\Omega_{b0}}{\Omega_{m0}}$$

$$\overset{(8.52)}{=\!=}\ \frac{3}{4}\times 1.6183\times\frac{\Omega_{b0}}{\Omega_{m0}} = 0.202\left(\frac{\Omega_{b0}}{0.045}\right)\left(\frac{\Omega_{m0}}{0.27}\right)^{-1} \tag{9.79a}$$

$$R_{b,dc} = \frac{R_{\text{dc}}}{R_{\text{eq}}}R_{b,\text{eq}} = 0.592\left(\frac{1+z_{\text{eq}}}{3200}\right)\left(\frac{1+z_{\text{dc}}}{1091}\right)^{-1} \tag{9.79b}$$

where we used $\rho_m(t_{\text{eq}}) = \rho_{\text{rad}}(t_{\text{eq}})$ and that both ratios ρ_b/ρ_m and $\rho_\gamma/\rho_{\text{rad}}$ are conserved numbers. Substituting Eqs. (9.79a, b) and (8.137) into Eq. (9.78), we obtain

$$\begin{aligned} d_s(t_{\text{dc}}) &= 147.6\text{Mpc} \simeq 105\,\text{h}^{-1}\,\text{Mpc}^{9)} \\ k_{\text{dc}} &= \frac{2\pi}{d_s(t_{\text{dc}})} = 0.06\,\text{h}^{-1}\,\text{Mpc}^{-1} \end{aligned} \tag{9.80}$$

which agrees with the BAO peak observed by the SDSS/BOSS survey in Figure 9.3(b).

9.1.7
Initial fluctuation

Harrison–Zeldovich Spectrum As stated in the previous section, the fluctuation is in the linear regime if the size (as converted to today's scale) is larger than $\lambda > 20$ Mpc. Then it is possible to trace its origin back in time to the initial distribution. When the universe expands in powers of the scale factor ($R \propto t^\alpha$, $\alpha = 2/3$ for matter and $\alpha = 1/2$ for radiation dominance), the horizon is given by $d_H(t) = t(1-\alpha)^{-1}$

[9] Ref. [760] quotes $d_s(t_{\text{dc}}) = 154.7$ Mpc. A fitting formula to obtain the first peak is given in [759].

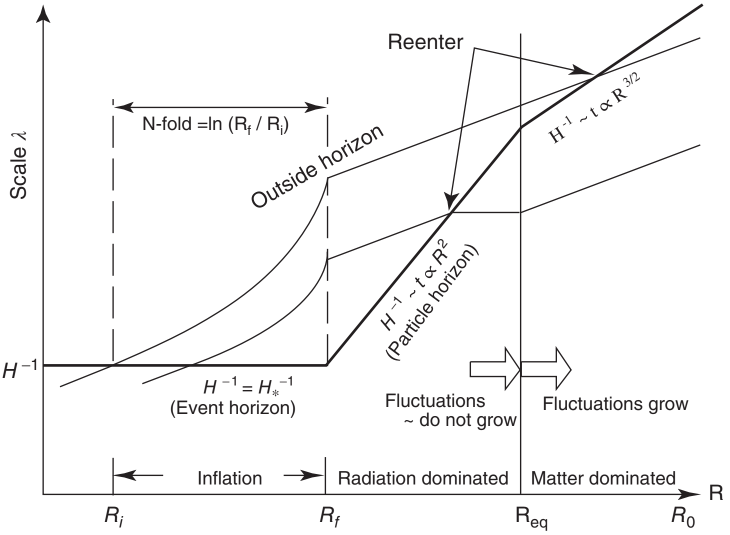

Reenter
N-fold =ln (R$_f$ / R$_i$)
Scale λ
Outside horizon
$H^{-1} \sim t \propto R^{3/2}$
$H^{-1} \sim t \propto R^2$
(Particle horizon)
H^{-1}
$H^{-1} = H_*^{-1}$
(Event horizon)
Fluctuations
~ do not grow
Fluctuations grow
Inflation
Radiation dominated
Matter dominated
R
R_i
R_f
R$_{eq}$
R_0

Figure 9.5 Illustration of the scale growth as compared to the Hubble scale. Fluctuations at scale $\lambda = H_*^{-1}$ crosses the horizon drawn in a thick line during the inflation era, stretches, and re-enters the horizon at later time. After re-entry into the horizon, the fluctuations that had entered before the matter-dominated era ($R < R_{\text{eq}}$) do not grow (or grow at most logarithmically) until the matter–radiation equality. Note that the scale of the abscissa for $R < R_f$ is changed to emphasize exponential growth of the inflation.

[see Eq. (8.77)]. Therefore, if one specifies a certain scale λ at the present time, their time dependence is given by

$$\lambda \propto t^{2/3}, \qquad d_H(t) \sim 3t \tag{9.81}$$

which means that, in the past and for sufficiently small t, the scale λ was larger than the horizon size and was outside the horizon. As the dynamical structure formation can be active only within the causal boundary, the fluctuation at the crossing of the horizon provides the initial value.

Conventionally, the initial power spectrum is assumed to obey a power law.

$$\delta_k^2 = Ak^n \tag{9.82}$$

The choice of $n = 1$ is referred to as the *Harrison–Zeldovich* isocurvature spectrum [761, 762]. To specify the initial value of the fluctuation, we need to know the initial time and the size of the fluctuation when it enters the horizon. For this, we need to discuss time evolution of super-horizon fluctuations.

At near or beyond the horizon size, the Newtonian equations are no longer valid. Even the concept of the density fluctuation becomes obscure. This is because the general relativity has the freedom of gauge (coordinate) transformation, and the density fluctuation is not a gauge-invariant quantity (see Appendix E). Super-horizon fluctuations do evolve, but, being separated beyond the reach of causality, the evolution is not induced by dynamics as a result of interactions but basically by kinematics. Therefore, it is to be considered as a ripple in the fabric of spacetime [125]. One has to rely on general relativity for detailed discussions. Rigorous treatments based on general relativity are beyond the scope of this book. However, simplified qualitative arguments are possible to understand the origin of the initial fluctuation.

As we are discussing deviations from the uniform isotropic universe, we need to consider perturbations to the Robertson–Walker metric. They appear as the gravitational potential Ψ and the space curvature Φ.

In the Newtonian gauge (see Eq. (9.10)), the curvature fluctuation Φ is equal to the negative of the gravitational potential Ψ for a perfect fluid which we are assuming. The density fluctuation is related to it by Eq. (9.22). It is derived nonrelativistically and has to be modified for relativistic super-horizon problems. However, the following arguments do not change if we use the nonrelativistic equation. From Eq. (9.22), one sees that the curvature fluctuation expressed as the power spectrum is proportional to the power spectrum of the density fluctuation.

$$\Delta_\Phi = \sqrt{<\Phi_k^2>} = \frac{4\pi G_N}{k^2} R^2 \rho_0 \Delta_k = \frac{3}{2}\left(\frac{RH}{k}\right)^2 \Delta_k \sim k^{\frac{n-1}{2}} \tag{9.83}$$

where we used the Friedmann equation $\left(H^2 \sim (8\pi/3) G_N \rho_0\right)$. Equation (9.83) means that the curvature fluctuation does not depend on the scale $1/k$ for $n = 1$; hence the $n = 1$ spectrum is also referred to as the *isocurvature spectrum*. It also means that the density fluctuation is equal to the curvature fluctuation (times a factor 2/3) at the horizon entrance ($1/k = 1/RH = d_H$). That is, the $n = 1$ spectrum has a characteristic feature that it has the same magnitude independent of the scale.

Now let us see how the fluctuation evolves after entering the horizon. We want to understand the observed k-dependence of the power spectrum qualitatively, which is depicted in Figure 9.6(a). It is derived from the combined analyses of CMB, the galaxy distributions depicted in Figure 9.1, Lyman α spectra, and so on. As the horizon is growing with time, the fluctuation of larger scale (smaller k) enters horizon at later time. For scales $k < k_{\text{eq}}$, where $k_{\text{eq}} = 1/d_H(t_{\text{eq}})$ is the inverse horizon scale at the time of matter–radiation equality, the universe has already entered the matter-dominated era. The fluctuations grow like $\sim R$ independent of k, and therefore it retains the original, that is, the initial k dependence $\delta_k^2 \sim k^n$.

On the other hand, for smaller scale ($k > 1/d_H(t_{\text{eq}})$), the fluctuation enters the horizon while the universe is still radiation-dominated and cannot grow (or grows at most logarithmically) until $t = t_{\text{eq}}$. Thereafter, fluctuations with $k > k_{\text{eq}}$ grow uniformly. Hence, the strength of the fluctuation should be roughly constant ($\Delta_k \sim$ constant) independent of the scale, that is

$$\delta_k^2 \sim \begin{cases} k & k < k_{\text{eq}} \\ k^{-3} & k > k_{\text{eq}}. \end{cases} \tag{9.84}$$

To find the position of k_{eq} in Figure 9.6(a), we have to express it in the comoving coordinate system.

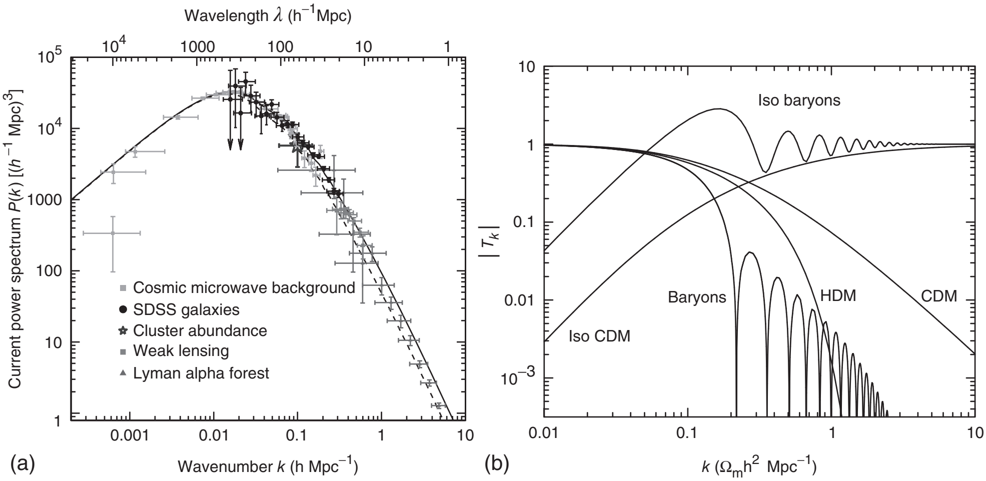


Figure 9.6 (a) Power spectrum (δ_k^2): data compared with theories (ΛCDM model with $\Omega_{m0} = 0.28, h = 0.72$, and $\Omega_{b0}/\Omega_{m0} = 0.16$). The dashed curve shows a suppression of small-scale power by about a factor of 2 due to a replacement of 7% of the CDM with a massive neutrino having $\sum_j m_j = 1$ eV. (Reproduced with permission of [763].) (b) Plot of the transfer functions for various adiabatic models, in which $T_k \to 1$ at small k. A number of possible matter contents are illustrated: pure baryons; pure CDM; pure HDM. (Reproduced with permission of [7, 764].)

$$1/k_{\text{eq}} = d_H(t_{\text{eq}}) = 2ct_{\text{eq}} \tag{9.85}$$

$$\begin{aligned} \to \quad d_H(t_{\text{eq}})|_{\text{comov}} &= 1/R_{\text{eq}}H(t_{\text{eq}}) = d_H(t_{\text{eq}})(1+z_{\text{eq}}) \simeq 57.4\,\text{h}^{-1}\text{Mpc}, \\ \text{or} \quad k_{\text{eq}} &\simeq 0.017\,\text{h Mpc}^{-1} \end{aligned} \tag{9.86}$$

This spectrum reproduces the qualitative features of the observed power spectrum in Figure 9.6(a). More quantitatively, one defines the transfer function $T(t_0, k)$ by

$$P(t_0, k) = T^2(t_0, k)P_{\text{init}}(k) \tag{9.87}$$

where $P_{\text{init}}(k)$ is the initial fluctuation which is generally taken as $P_{\text{init}}(k) = Ak^n$. The transfer function has to be calculated numerically for each k using the density fluctuation evolution equation based on the Boltzmann equation. It decreases for $k > k_{\text{eq}}$ as discussed. Some representative functions for differing models are given in Figure 9.6(b).

As the Jeans length of the neutrino is ~100 Mpc, the HDM spectrum has a cutoff at $k \sim 1/100 \sim 0.1\,\Omega_m h^2 \text{Mpc}^{-1}$. The dashed curve in Figure 9.6(a) depicts such an effect. For the CDM, there is no such cutoff, and the spectrum extends to large k.

In comparing cosmic models with observations, one sets the initial conditions at some time $t < t_{\text{eq}}$, assumes models (ΛCDM, HDM etc.) and Gaussian distribution for the phase of the fluctuation, chooses parameters for Ω_{m0}, Ω_{rad0}, Ω_Λ, Ω_K, and so on, and carries out simulation calculations. Figure 9.6(a) compares the ΛCDM model with observed data.

9.1.8
Effects of Neutrino Mass

Although the neutrino is no longer considered as a main component of the dark matter, it can still affect the large-scale structure of the universe if its mass is on the order of electron volts. The scale of the neutrino's Jeans wavelength (see Eq. (9.52)) is approximately the horizon size at the time of matter–radiation equality. Consequently, damping due to the neutrino's attraction introduces a cutoff at $k > k_{\text{eq}} = 2\pi/\lambda_{\text{eq}}$ to the power spectrum, which can be used to determine the neutrino mass.

Several curves at $k > 10^{-2}$ (h/Mpc) in Figures 9.7 show how the neutrino mass can affect the power spectrum of the universe. Recent analyses take into account the neutrino mass effect on both CMB and large-scale structure and obtain upper limits of the neutrino mass typically on the order [120, 121, 766]

$$\sum m_\nu \lesssim 0.1{\sim}0.3 \text{ eV}. \tag{9.88}$$

9.1.9
Primordial Fluctuation

The initial condition we prepared to solve the evolution of the density fluctuation is most likely scale-free. Its simplicity suggests a more fundamental origin. The inflation models can provide a scenario for the origin of the initial fluctuation. Because of its super-horizon nature, we conjecture that it was generated in the pre-inflation era. An attractive idea is to consider that its origin is the quantum mechanical fluctuation of the inflaton field. It is the ultimate fluctuation, which

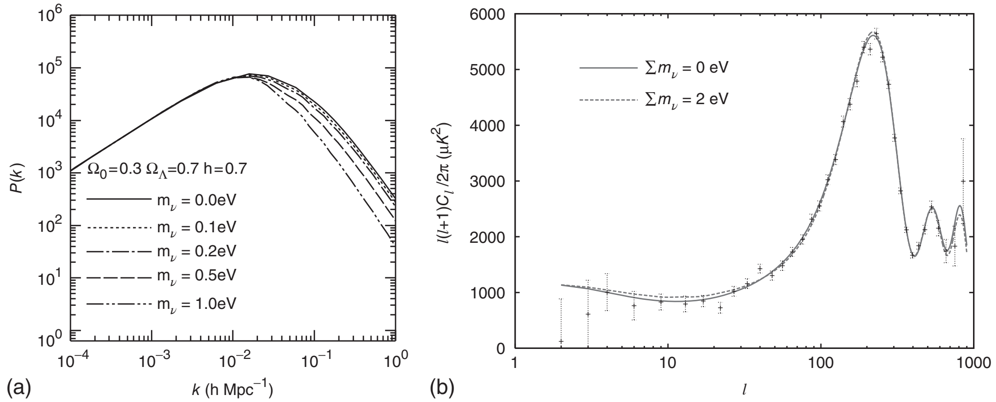


Figure 9.7 (a) Power spectrum predicted in the ΛCDM model including massive neutrinos with degenerate mass specified in the legend. The top curve represents the case of massless neutrinos. (b) CMB multipoles for the ΛCDM solution with massive neutrinos that is allowed just at a 95% confidence ($\sum m_\nu = 2.0$ eV, dashed curve) are compared with those for the best ΛCDM solution with massless neutrinos (solid curve). The actual constraint from the CMB data is more severe. The data are from WMAP. (Reproduced with permission of [765].)

we refer to as the *primordial fluctuation*. It starts out as micro fluctuations, crosses the (event) horizon, stretched by the inflation, and re-enters the horizon at the post-inflation era providing the initial fluctuation (see Figure 9.5).

Let us calculate the quantum fluctuation generated by the inflaton field. As during the inflation epoch the potential is nearly constant (slow roll: see arguments after Eq. (8.151)), it means that the inflaton can be considered as an almost massless field ($m^2 = \partial^2 V/\partial\phi^2 \simeq 0$). Let us consider a free-field Lagrangian of the massless inflaton in the expanding universe

$$\int d^3x\, R^3 \mathscr{L} = \frac{1}{2}\int d^3x\, R^3 \partial_\mu \phi \partial^\mu \phi = \sum_{\boldsymbol{k}} \frac{R^3}{2}\left(\dot{\phi}_{\boldsymbol{k}}^2 - \frac{\boldsymbol{k}^2}{R^2}\phi_{\boldsymbol{k}}^2\right) \tag{9.89}$$

where $\phi_{\boldsymbol{k}}$ is the Fourier transform of the inflaton field. The equation of motion in the comoving coordinate system for each mode is given by

$$\ddot{\phi}_{\boldsymbol{k}} + 3H\dot{\phi}_{\boldsymbol{k}} + \frac{|\boldsymbol{k}|^2}{R^2}\phi_{\boldsymbol{k}} = 0 \tag{9.90}$$

where the Hubble expansion rate during the inflation era is kept approximately constant $H = H_* = \sqrt{\Lambda/3}$. [see Eq.(8.56)] This is equivalent to the equation of a harmonic oscillator having mass $m = R^3$ and frequency $\omega = |\boldsymbol{k}|/R = k/R$ moving under the frictional force $3H$. The friction is generated by the expansion of the universe. For $k/R \ll H$ or at a scale larger than the horizon ($d_H = 1/RH$), the friction prevents any motion, that is, the field is frozen and the fluctuation remains constant. It should be so because no interaction can affect the motion of fluctuations beyond the horizon. For scales smaller than the horizon, the friction term can be neglected and the field oscillates. The inflaton field amplitude fluctuation translates into the position uncertainty in the harmonic oscillator problem. It is a basic problem of quantum mechanics. We know that the fluctuation is given by $<(\delta x)^2>= 1/2m\omega$ with a Gaussian shape. Translated back to the inflaton field fluctuation, we obtain

$$|\delta\phi_{\boldsymbol{k}}|^2 = \frac{1}{2R^2 k}. \tag{9.91}$$

During the inflation, the horizon is kept almost constant. The scale at the time when the fluctuation goes out of the (event) horizon is given by $R/k = 1/H_*$. From Eq. (9.91), the fluctuation outside the horizon is given by

$$|\delta\phi_{\boldsymbol{k}}|^2 \sim \frac{H_*^2}{2k^3} \tag{9.92}$$

where H_* is the Hubble constant at the time of event horizon crossing. As the magnitude of the fluctuation at scale k (i.e., the power spectrum) is expressed by Eq. (9.54), we define the inflaton field fluctuation at scale $r = 1/k$ as

$$(\delta\phi)^2 \simeq \left. <\delta\phi(\boldsymbol{r})^2> \right|_{r=1/k} \equiv \Delta_\phi^2 = \frac{k^3|\delta_{\phi_k}|^2}{2\pi^2} = \left(\frac{H_*}{2\pi}\right)^2. \tag{9.93}$$

The connection between the inflaton fluctuation in the pre-inflation era and the density fluctuation in the post-inflation era has to be treated carefully using the Einstein equation. Intuitively, one can connect the two as follows: If the inflaton field has spatial fluctuation, it affects the time of inflation disappearance (t_2) through Eq. (8.156), that is, the time of reheating when the inflaton energy is converted back to matter or the radiation energy fluctuates.

$$\delta t = \frac{\delta\phi}{\dot{\phi}} \tag{9.94}$$

If the reheating commences at different times while the universe is expanding, the time fluctuation is converted to the energy density fluctuation. As $\rho \propto 1/R^3$ and $H \simeq H_*$ =constant

$$\delta = \frac{\delta\rho}{\rho} = 3\frac{\delta R}{R} = 3H\delta t = 3H_*\frac{\delta\phi}{\dot{\phi}}. \tag{9.95}$$

Substituting Eq. (9.93) into Eq. (9.95), we have for the density fluctuation at scale k

$$\delta \simeq H_*\frac{\delta\phi}{\dot{\phi}} = \frac{H_*^2}{2\pi\dot{\phi}} \equiv \delta_H. \tag{9.96}$$

As both H_* and $\dot{\phi}$ are nearly constant during the inflation epoch, one concludes that the super-horizon fluctuation is nearly scale-independent. As the fluctuation is frozen outside the horizon, we obtain the scale-independent initial fluctuation as it enters the horizon. A small scale dependence (i.e., finite value of $n-1$) is produced by the nonzero slope of the slow rolling potential.

Tilt The density fluctuation at the horizon crossing is given by δ_H^2 in Eq. (9.96). The shift of the index from the $n = 1$ power spectrum is therefore given by

$$\text{tilt} \equiv n - 1 = -\frac{d\ln\delta_H^2}{d\ln k}. \tag{9.97}$$

As the wavenumber at crossing is given by $k = RH$ and as we can assume H is almost constant during the inflationary evolution, we can replace $d\ln k$ by $d\ln R$.

$$\frac{d}{d\ln k} = R\frac{d}{dR} = R\frac{\dot{\phi}}{\dot{R}}\frac{d}{d\phi} = \frac{\dot{\phi}}{H}\frac{d}{d\phi} = -\frac{1}{8\pi G_N}\frac{V'}{V}\frac{d}{d\phi} \tag{9.98}$$

where we have used Eq. (8.153) and the Friedmann equation Eq. (8.10). Noting

$$\delta_H^2 = \left(\frac{H^2}{2\pi\dot{\phi}}\right)^2 \overset{(8.153)}{=\!=} \frac{128\pi G_N^3}{3}\frac{V^3}{V'^2} = \frac{1}{24\pi^2}\frac{1}{\tilde{M}_{Pl}^4}\frac{V}{\epsilon} \tag{9.99}$$

and using definitions of the slow rolling parameters ε and η in Eqs. (8.154), we have

$$1 - n = 6\varepsilon - 2\eta. \tag{9.100}$$

Deviation of the tilt from unity has an effect on the CMB spectrum described in the next section. As the slow rolling parameters can be calculated for a given inflaton potential and $1-n$ can be obtained from measurements of CMB spectrum (see Table 9.1), Eq. (9.100) can be used to choose the correct inflation model. WMAP and Planck [705, 767] report a value of $n = 0.961 \pm 0.0054$.

In summary, scale-free fluctuation, the small tilt, and Gaussianity of the CMB distribution strongly suggest the inflationary origin of the fluctuation.

Table 9.1 Summary of the cosmological parameters of ΛCDM Model. (Reproduced with permission of [123, 674, 705].)

Parameter	WMAP + BAO + SN	Planck + WP + high L + BAO (68% CL)
σ_8	0.812 ± 0.026	0.826 ± 0.012
H_0	$70.5 \pm 1.3\,\mathrm{km\,s^{-1}Mpc^{-1}}$	$67.80 \pm 0.77\ \mathrm{km\ s^{-1}Mpc^{-1}}$
$\Omega_{b0}h^2$	0.0227 ± 0.0006	0.02214 ± 0.00024
Ω_Λ	0.726 ± 0.015	0.692 ± 0.010
$\Omega_{m0}h^2$	0.136 ± 0.004	0.1426 ± 0.0025
$\Omega_{\mathrm{CDM}}h^2$	0.1138 ± 0.0045	0.1187 ± 0.0017
z^f_{reion}	10.9 ± 1.4	11.3 ± 1.1
t_0	13.72 ± 0.12 Gyr	13.798 ± 0.037 Gyrs
n_s	0.960 ± 0.013	0.9608 ± 0.0054
τ	0.084 ± 0.016	0.092 ± 0.013
w	$-1^{+0.12}_{-0.14}$	$-1.13^{+0.23}_{-0.25}$ (95% CL)
Ω_k	$0^{+0.0081}_{-0.0179}$	$-0.0005^{+0.0065}_{-0.0066}$ (95% CL)

Primordial Gravitational Wave Gravitational waves are an inevitable consequence of all inflationary models [768–770] because vacuum fluctuations are created for all the existing fields including the tensor metric field. They cause anisotropy and polarization of the CMB. From the Einstein equation [Eq. (8.1)] and decomposition of metric tensor into scalar, vector and tensor components, we find that in the linear scheme, two components of the gravitational waves $h_a = h_+, h_x$ behave as a free, massless scalar field with the normalization factor $\sqrt{16\pi G_N}$ (see Appendix Section E.4). That is, the amplitude h_a is suppressed by the Planck mass. As the scalar field fluctuation is given by Eq. (9.93), we immediately obtain the power spectrum for the gravity field fluctuation.

$$\Delta^2_{gw}(k) = 16\pi G_N \Delta^2_\phi(k) = \frac{1}{2\pi^2}\left(\frac{H}{\tilde{M}_{\mathrm{Pl}}}\right)^2_{k=HR} \tag{9.101}$$

As H is nearly constant during the inflation, Δ^2_{gw} is nearly scale-independent just like the primordial inflaton fluctuation. The amplitude remains constant after exiting and re-entering the horizon. However, the amplitude with scale smaller than the horizon decays after reentry (see Section E.4). The relative strength of the tensor mode to the scalar mode can be obtained as follows:

$$\Delta^2_{gw}(k) = 2 \times 2\left(\frac{H}{2\pi\tilde{M}_{\mathrm{Pl}}}\right)^2_{k=HR} \simeq \frac{1}{3\pi^2}\frac{V}{\tilde{M}^4_{\mathrm{Pl}}} \tag{9.102}$$

where V is the inflation potential and the pre-factor 2 takes the two polarization degrees of freedom into account. The strength of the scalar amplitude is given by Eq. (9.83). Using Eq. (9.99)

$$\Delta^2_\Phi = \frac{9}{4}\left(\frac{RH}{k}\right)^4_{k=RH}\Delta^2_k \simeq \frac{9}{4}\frac{128\pi G_N^3}{3}\frac{V^3}{V'^2} = \frac{9}{4}\frac{1}{24\pi^2}\frac{1}{\tilde{M}^4_{Pl}}\frac{V}{\epsilon}. \tag{9.103}$$

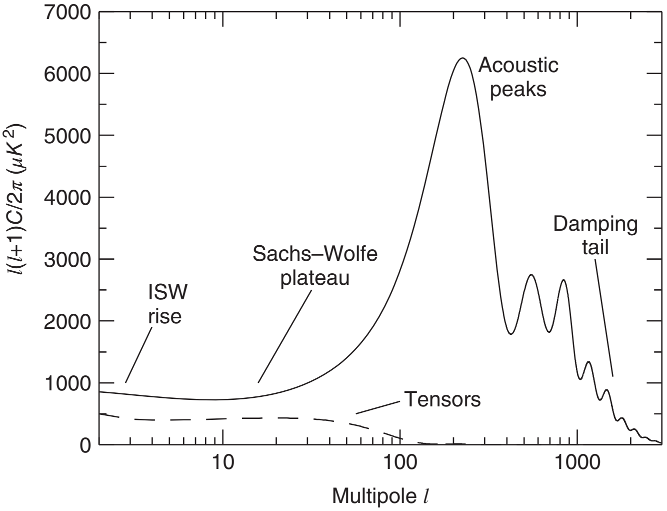


Figure 9.8 Contributions to CMB anisotropies from scalar and tensor (gravitational wave) perturbations generated during inflation. The concordance ΛCDM model has been assumed with exactly scale-invariant fluctuations. The amplitudes of the tensor and scalar power spectra are arbitrary. The tensor amplitude is approximately flat at the super-horizon scale ($l \lesssim 100$) and damped at sub-horizon scale. (Reproduced with permission of [8].)

Then, the tensor-to-scalar ratio $r = \Delta_{gw}^2/\Delta_\Phi^2$ becomes $\sim 3.5\epsilon$, where ϵ is the slow rolling parameter given by Eq. (8.154). A more rigorous treatment gives [771]

$$r = \frac{\Delta_{gw}^2}{\Delta_\Phi^2} \simeq 16\epsilon, \qquad \epsilon = \frac{\tilde{M}_{\rm Pl}^2}{2}\left(\frac{V'}{V}\right)^2. \tag{9.104}$$

Figure 9.8 illustrates how the primordial gravitational waves show up in the CMB power spectrum.

The ratio r is a function of k. However, as Figure 9.8 shows, we may consider it as approximately a constant for $l \lesssim 100$. As the larger scale enters the horizon later and $k^{-1} \simeq 2/(Hl)$, the relevant range $2 < l \lesssim 100$ corresponds to $\Delta \ln k \sim 4$. Consequently, the universe expands at least 4 e-fold during the period while the relevant components of the gravitational wave pass the horizon.

Combining equations (8.156) and (9.104), we can constrain the value of r.

$$\Delta N \simeq \frac{1}{\tilde{M}_{\rm Pl}^2}\frac{V}{V'}\Delta\phi = \frac{1}{\sqrt{2\epsilon}}\frac{\Delta\phi}{\tilde{M}_{\rm Pl}} \tag{9.105}$$

$$r = 16\epsilon = \frac{8}{\Delta N^2}\left(\frac{\Delta\phi}{\tilde{M}_{\rm Pl}}\right)^2 < \frac{8}{4^2}\left(\frac{\Delta\phi}{\tilde{M}_{\rm Pl}}\right)^2 = 0.5\left(\frac{\Delta\phi}{\tilde{M}_{\rm Pl}}\right)^2 \tag{9.106}$$

which is referred to as the Lyth bound [772]. However, considering the observed near scale independence of the spectral index [767], a more plausible value can be obtained by extending ΔN to the entire observable sky. Therefore we set $\Delta N \simeq 60$

which is the value necessary to solve the horizon problem (see Section 8.8.2). Then

$$r \approx 2 \times 10^{-3} \left(\frac{\Delta\phi}{\tilde{M}_{\rm Pl}} \right)^2 \tag{9.107}$$

This means that r is measurable only for large field ($\Delta\phi \gg \tilde{M}_{\rm Pl}$) models.

For the power potential $V \sim \phi^\alpha, \epsilon = (\alpha^2/2)(\tilde{M}_{\rm Pl}/\phi)^2, \eta = \alpha(\alpha - 1)(\tilde{M}_{\rm Pl}/\phi)^2$ and $\phi = \sqrt{2\alpha N}\tilde{M}_{\rm Pl}$ can be obtained from Eq. (8.156). Then Eq. (9.100) gives

$$r = 8(1 - n) - \frac{8}{N} \simeq 0.133, \ \epsilon = \frac{\alpha}{4N} \simeq 0.0083 \text{ for } \alpha = 2, N = 60 \tag{9.108}$$

In this model, the gravitational waves are big enough to be eventually observable.

There is an indication that the value of r is much smaller than that of the naive polynomial models as given in Eq. (9.108). Figure 9.9 shows the upper limits of r based on the combined analysis of the CMB power spectra and BAO data. The Planck team obtained the spectral index $n = 0.9603 \pm 0.0073$ and an upper limit on r [767].

$$r < 0.11 \quad 95\% \text{ CL.} \tag{9.109}$$

As the value of ϵ is highly model-dependent, r is treated as an adjustable parameter in the CMB anisotropy analysis, which is our next subject.

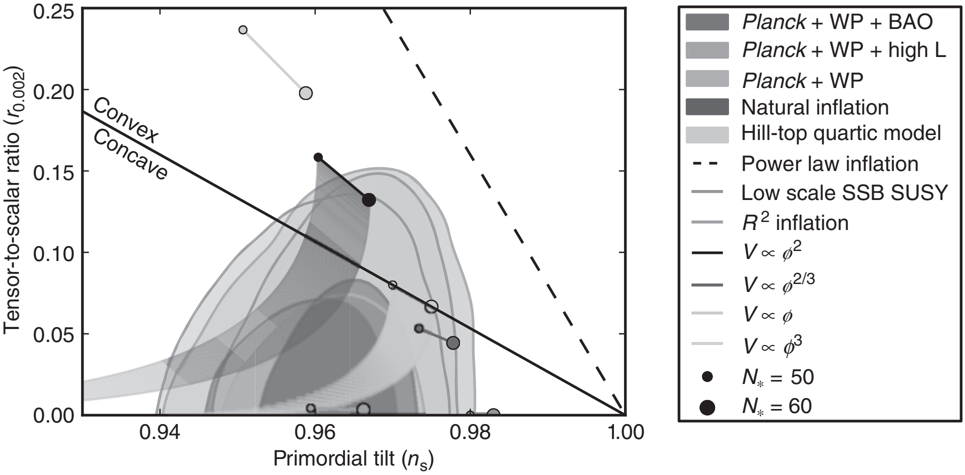


Figure 9.9 Marginalized 68% and 95% confidence levels for n_s and r from Planck + WMAP and BAO data, compared to the theoretical predictions of selected inflationary models. (Reproduced with permission of ESA and [767].) (Please find a color version of this figure on the color plates.)

9.2
CMB Anisotropy

9.2.1
Overview

Existence of the CMB is the strongest evidence in support of the hot Big Bang cosmology. Its discovery by Penzias and Wilson in 1967 [773] may be heralded as the first revolution in cosmology. The early universe was a furnace of all species of particles in thermal equilibrium. After about 3 minutes past the Big Bang (i.e., after the nucleosynthesis), the universe was in a plasma state filled with a very large number of photons mixed with a small number of protons, helium nuclei, and electrons. Around $t = 380,000\,(z \simeq 1090)$ years, the temperature dropped down to much lower than that of the hydrogen ionization energy, prompting electrons and protons to form neutral atoms (recombination). When almost all of the charged particles disappear, the photon is no longer scattered, travels freely, and reaches us, which is observed as the CMB. It is observed as the electromagnetic wave noise coming uniformly from all over the sky.[10] The CMB photons are radiated from a thin layer of the cosmic sphere located at $z = 1090$, which is referred to as the *surface of last scattering*.

The temperature distribution of the CMB is uniform over all the sky. When the tiny but finite anisotropy[11] was detected in 1992 by the COBE satellite [774], it excited the astronomical community so much that it may be called the "second revolution in cosmology".

Its spectrum is shown in Figure 9.10(a), which follows that of the blackbody radiation within an error 10^{-5} and its temperature is determined very accurately ($T_0 = 2.725 \pm 0.001$ K [7, 778]). As stated before, the smallness of the anisotropy is the strongest supporting evidence for the cosmological principle and also provides a theoretical reasoning for the existence of the dark matter. Precision data by WMAP established the ΛCDM model and determined various cosmological parameters to an unprecedented precision. Among others, the CMB power spectrum analysis gave the observational evidence for the inflation model.

We will describe below how to extract information from the CMB spectrum. The temperature anisotropy $\delta T/T(\hat{\boldsymbol{r}})$ at a designated direction $\hat{\boldsymbol{n}} = (\theta, \phi)$ can be expanded as

$$\frac{\delta T}{T_0}(\hat{\boldsymbol{n}}) = \boldsymbol{\beta} \cdot \hat{\boldsymbol{r}} + \sum_{l=2} \sum_{m} a_{lm} Y_{lm}(\theta, \phi). \tag{9.110}$$

$T_0 = 2.725 \pm 0.001\,K$ is the average value of the temperature. If the anisotropy is statistically isotropic, it does not depend on ϕ.

$$< a_{lm} a^*_{l',m'} >= \delta_{ll'} \delta_{mm'} \; < |a_l|^2 > \tag{9.111}$$

---

[10] Any radiowave antenna picks the CMB waves. With advanced technologies, we may be able to sort it out from the television noises.

[11] Spatial irregularities in the power spectrum of CMB is referred to as *anisotropy* rather than fluctuation as is used for matter particles.

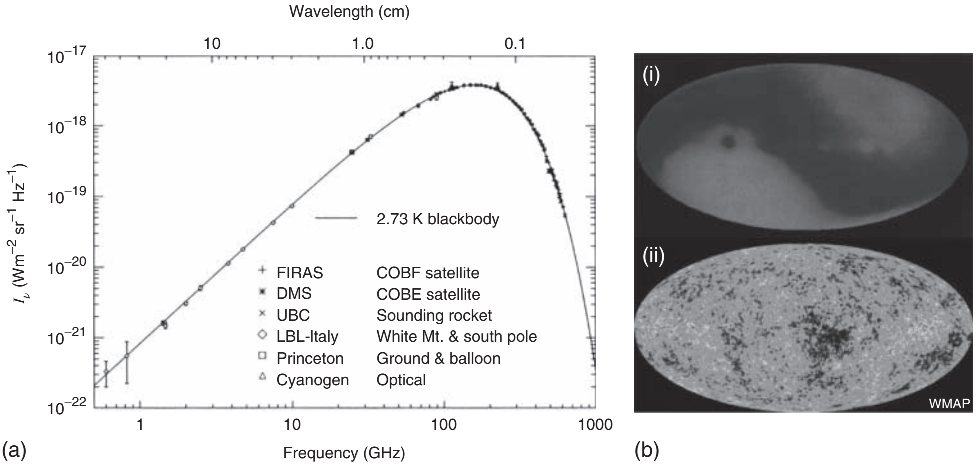


Figure 9.10 (a) CMB spectrum agreeing with the Planck's blackbody radiation formula with error (smaller than the thickness of the line) smaller than 10^{-5}. (Reproduced with permission of [775].) (b) Color-coded anisotropy of CMB. (i): Up to the accuracy of 10^{-3} only a dipole component is observed. [776] (ii): After subtraction of the dipole component, finite anisotropy of $\sim 10^{-5}$ is observed. Figure courtesy of NASA [777]. Figure (a) was obtained by analyzing this anisotropy map. (Please find a color version of this figure on the color plates.)

According to observations, the first term in Eq. (9.110) is overwhelmingly larger than the others:

$$\begin{aligned} T_1 &\equiv T_0\,\beta = 3.61 \pm 0.1 \times 10^{-3}\,\mathrm{K} \\ v &= c\beta = 369.19 \pm 19\,\mathrm{km}\ \ [779]. \end{aligned} \tag{9.112}$$

It is obtained from the Planck formula Eq. (8.38) by a Lorentz transformation with velocity $\boldsymbol{\beta}$. It is observed as the dipole component of the spherical distribution (see Figure 9.10b, i). It is considered as the result of the Doppler shift due to motion of the Earth with respect to the comoving coordinate system. Therefore, it does not contain cosmological information and is subtracted for the study of the CMB.

The relativistic Doppler shift is given by

$$\nu_{\mathrm{obs}} = \nu_{\mathrm{rest}} \frac{\sqrt{1 - u^2/c^2}}{1 + (u/c)\cos\theta} = \frac{\nu_{\mathrm{rest}}/\gamma}{1 + v_r/c} \tag{9.113}$$

where ν is frequency of the light, u is relative velocity of the light source and an observer, and θ is the angle between the emitted light and the source velocity. $\gamma = 1/\sqrt{1-\beta^2}$ is the time dilatation effect. Equation (9.113) translates to temperature change of the Planck's formula

$$T_{\mathrm{moving}} = \frac{T_{\mathrm{rest}}/\gamma}{1 - (u/c)\cos\theta} \simeq T_{\mathrm{rest}}(1 + \boldsymbol{\beta} \cdot \hat{\boldsymbol{n}}) \tag{9.114}$$

which reproduces the first term of Eq. (9.110). The CMB anisotropy map after the dipole subtraction obtained by the WMAP is shown in Figure 9.10b, ii. The most recent map obtained by the Planck satellite is shown in Figure 9.11.

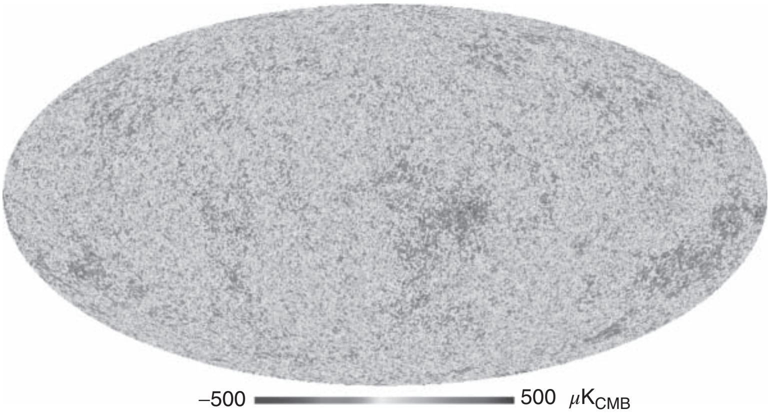


Figure 9.11 Cosmic microwave background (CMB) radiation anisotropy map across the sky obtained by the Planck satellite. The dipole component is subtracted. The intensity of the anisotropy is $O(10^{-5})$. Figure by ESA (Reproduced with permission of [674, 780].). (Please find a color version of this figure on the color plates.)

Analysis of the CMB anisotropy obtained by the WMAP/Planck occupies the rest of the discussions in this chapter.

A convenient variable that is connected with the measurable quantities and depends on the statistical distribution of the temperature anisotropy is the two-point function defined by

$$\begin{aligned} C(\theta) &= \left\langle \frac{\delta T}{T_0}(\boldsymbol{n}_1)\frac{\delta T}{T_0}(\boldsymbol{n}_2) \right\rangle \\ \cos\theta &= \boldsymbol{n}_1 \cdot \boldsymbol{n}_2. \end{aligned} \tag{9.115}$$

Observationally, what is often measured is the rms (root-mean-square) temperature difference between two antennas separated by an angle θ averaged over some substantial fraction of the sky: $\delta T/T \equiv <(T_1 - T_2)^2>^{\frac{1}{2}} /T$. It is related to the two-point function by

$$\left(\frac{\delta T}{T}\right)^2 = 2\big[C(\theta) - C(0)\big]. \tag{9.116}$$

As $\delta T/T$ is real, we have

$$C(\theta) = \sum_{l_1,m_1}\sum_{l_2,m_2} < a_{l_1,m_1} a^*_{l_2,m_2} > Y_{l_1,m_1}(\theta_1,\phi_1)Y^*_{l_2,m_2}(\theta_2,\phi_2). \tag{9.117}$$

Using Eq. (9.111) and defining the angular power spectrum by

$$C_l = \frac{1}{2l+1}\sum_{m=-l}^{+l} < |a_{l,m}|^2 > \tag{9.118}$$

one obtains

$$C(\theta) = \sum_l \frac{2l+1}{4\pi} C_l P_l(\theta) \tag{9.119}$$

where $P_l(\theta)$ is the Legendre function of order l and we have used the addition theorem of spherical harmonics. Therefore, the two-point correlation function gives information on the angular power spectrum. As the number of zero-crossings of the Legendre function P_l between $-1 < \cos\theta < 1$ is given by l, the quantity $(2l+1)C_l$ expresses the intensity of the anisotropy sliced at the angle of view $\theta \sim 180°/l$.

If the temperature anisotropy is Gaussian as predicted by the inflation models, the statistical distribution of each multipoles has zero mean $<a_{lm}>= 0$ and is fully characterized by only one parameter, namely the variance C_l. C_l can be predicted using theoretical models and compared with the experimentally measured value $\hat{C}_l$. Note that if the value of $\hat{C}_l$ is measured with infinite precision experimentally, there still is an intrinsic uncertainty, referred to as the *cosmic variance*. This is because, for small l, the number of samples we can extract is limited. For instance, for $l = 0$, there is only one sample in the whole sky and, unless one goes to the neighboring universe beyond the horizon, one cannot obtain another sample. If the multipoles are of random Gaussian types, $\sum_m |\hat{a}_{lm}|^2/C_l$ has a χ^2 distribution with $2l+1$ degrees of freedom and its variance is $2(2l+1)$. It is an inherent uncertainty, and even a perfect measurement cannot reduce it. Figure 9.12 shows the observed data fitted with the ΛCDM model [123,781]. An oscillation pattern of the anisotropy is quite conspicuous.

The origin of anisotropies in the cosmic microwave background can be traced to several sources: baryon–photon acoustic oscillation, gravitational red shift (Sachs–Wolfe (SW) effect), and Doppler shifts.

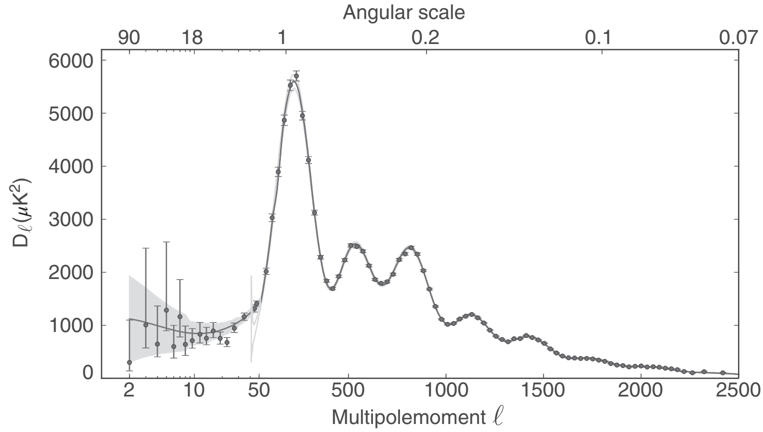


Figure 9.12 Temperature angular power spectrum of the primary CMB from Planck, showing a precise measurement of seven acoustic peaks, which are well fitted by a simple six-parameter ΛCDM theoretical model. The shaded area around the best-fit curve represents cosmic variance, including the sky cut used. The error bars on individual points also include cosmic variance. The horizontal axis is logarithmic up to $l = 50$, and linear beyond. The vertical scale is $l(l+1)C_l/2\pi$. (Reproduced with permission of ESA and [674,782].).

9.2.2
Sachs–Wolfe Effect

At the angle scale larger than ∼1°, the transverse size of the surface of the last scattering exceeds the horizon scale at the time of the last scattering, and hence the anisotropy is acausal. This means that the anisotropy is of a different origin from acoustic oscillation. Before the decoupling, the photons are in the gravitational potential created by the baryons and the dark matter. The photons have to crawl up the potential to reach us. It induces a gravitational red-shift effect which reflects the potential fluctuation. The effect is referred to as the *SW effect* [783]. Note that the observer is also in the gravitational potential. However, it affects equally over the entire temperature range and may enhance the average, but the anisotropy is not affected. The effect is scale-independent, that is, $\delta T/T\sim$ constant as a function of l.

If the photon is in the gravitational potential Ψ, its total energy outside the potential is $E = \text{KE} + \Psi$, where KE is the kinetic energy. As $\Psi < 0$ in the overdense region in spacetime, the photon temperature outside the potential is given by $\delta T_{\text{out}} = \delta T_{\text{in}} + \Psi$. If the photon reaches us at $T = T_0$, we have

$$\left(\frac{\delta T}{T}\right)_0 = \left(\frac{\delta T}{T}\right)_{\text{dc}} + \Psi_{\text{dc}} \tag{9.120}$$

where $\Psi_{\text{dc}} = \Psi(t = t_{\text{dc}})$. As the temperature goes down as the universe expands so that $TR =$ constant, we have $\delta T/T = -\delta R/R$. The scale changes with time as $R(t) \propto t^{2/3}$ in the matter-dominated era, which means

$$\frac{\delta R}{R} = \frac{2}{3}\frac{\delta t}{t}. \tag{9.121}$$

As the proper time in the gravitational potential changes to $d\tau = \sqrt{1 + 2\Psi}dt \simeq (1 + \Psi)dt$ (see Eq. (9.10)) so that

$$\frac{\delta t}{t} \simeq \Psi \tag{9.122}$$

then the temperature anisotropy at the time of decoupling becomes

$$\left(\frac{\delta T}{T}\right)_{\text{dc}} = -\frac{\delta R}{R} = -\frac{2}{3}\frac{\delta t}{t} = -\frac{2}{3}\Psi_{\text{dc}} \tag{9.123}$$

which when inserted into Eq. (9.120) becomes

$$\left(\frac{\delta T}{T}\right)_0 = \frac{1}{3}\Psi_{\text{dc}}. \tag{9.124}$$

Thus the CMB anisotropy at an angle of view larger than 1 degree tells us the gravitational potential at the time of last scattering, which is the primordial fluctuation inherited from the time of inflation because it is the anisotropy at the scale beyond the causal range.

In addition to the potential at the surface of scattering, light encounters various gravitational fields in its path to the Earth and the net effect on $\delta T/T$ is given by $2\int_{t_s}^{t_o} \dot{\Psi}\, dt$. The factor 2 comes from additional contribution of the curvature perturbation $-\Phi = \Psi$. This is referred to as the *integrated Sachs–Wolfe* (ISW) effect [784]. If the universe is matter-dominated, the potential stays constant[12] and the effect does not appear. This is because, when the potential is static (apart from expansion), the red-shift due to climbing the potential well and the blue shift due to sliding down the well compensate each other and the net effect vanishes. However, if the potential is changing with time, the compensation is not complete and a space-dependent red or blue shift is observed. There are two major causes.

One referred to as the *early ISW* is induced because, at the time of decoupling, the transition from the radiation dominance to the matter dominance is not complete. Especially, when the wavelength of the anisotropy is near the acoustic horizon, its growth is hindered and the potential decreases. This effect is sizable at the scale of the acoustic horizon immediately after the decoupling.

Another is referred to as the *late ISW*, which occurs in the later stage of cosmological development when the universe becomes Λ-dominated and the strength of the potential decreases. Therefore, light that comes through the overdense region is more blue-shifted while that passing through voids where the potential is higher is red-shifted. That is, the late ISW effect is related to the large-scale structure of the universe and can be used to investigate the effect of the dark energy.

9.2.3
Acoustic Oscillations

Assuming adiabatic fluctuations, the baryon–photon fluid before the decoupling oscillates as acoustic waves because of the photon pressure [Eq. (9.30)]. The gravity provides the attractive force among the baryons while the photon produces pressure. The pressure works effectively on electrons, but electrons and baryons are tightly coupled and the baryons respond to the photon pressure rapidly. As both the entropy and the baryon number are conserved, the tight coupling of the photons and baryons constrains the entropy $s \propto n_\gamma \propto \rho_\gamma^{3/4}$ of the photon per baryon to be a conserved number. Then

$$\frac{s}{n_b} \propto \frac{n_\gamma}{n_b} \propto \frac{\rho_\gamma^{\frac{3}{4}}}{\rho_b} = \text{constant} \quad \rightarrow \quad \delta_b = \frac{\delta\rho_b}{\rho_b} = \frac{3}{4}\frac{\delta\rho_\gamma}{\rho_\gamma} = \frac{3}{4}\delta_\gamma. \tag{9.126}$$

The temperature anisotropy due to the photon–baryon acoustic oscillation is related to the baryon and the photon density fluctuation by

$$\Theta \equiv \frac{\delta T}{T} = \frac{\delta_\gamma}{4} = \frac{\delta_b}{3}. \tag{9.127}$$

---

[12] $\nabla^2\Psi = 4\pi G_N \rho_0 R^2 \delta \sim R^{-3} \cdot R^2 \cdot R \sim \text{constant} \qquad (9.125)$

The baryon–photon fluid has the sound velocity as given by Eq. (9.74)

$$c_s^2 = \left(\frac{\delta P}{\delta \rho}\right)_{\text{adiabatic}} = \frac{1}{3}\frac{1}{1+R_b}, \qquad R_b = \frac{3}{4}\frac{\rho_b}{\rho_\gamma}. \tag{9.128}$$

Although the exact equation of CMB anisotropy has to be calculated using the Einstein equation, we see from Eq. (9.127) that qualitative features can be obtained from the Newtonian treatment of the baryon density fluctuation. The baryon fluctuation can be solved using Eq. (9.30). By using this equation, we are simplifying the problem that the universe is always matter-dominated. Using Eq. (9.22) and (9.128), Eq. (9.30) can be expressed as

$$\ddot{\delta}_b(t,\boldsymbol{k}) + 2H(t)\dot{\delta}(t,\boldsymbol{k}) + \frac{k^2 c_s^2}{R^2(t)}\left[\delta_b(t,\boldsymbol{k}) + 3(1+R_b)\Psi(t,\boldsymbol{k})\right] = 0. \tag{9.129}$$ [13]

This is an equation for forced oscillation with friction. If the background energy density (i.e., total density including the dark matter and neutrinos) is much larger than the baryon–photon compound, the potential is insensitive to δ_b or δ_γ and can be considered as an external field. For $H \ll kc_s/R$ (i.e., the oscillation period is much less than the Hubble time), the friction can be neglected and Eq. (9.129) reduces to an equation for a harmonic oscillator. R_b is a growing function of the scale parameter ($R_b \sim \rho_b/\rho_\gamma \sim R^{-3}/R^{-4} \sim R$) but changes appreciably only for time span of the Hubble time scale. So, R_b can also be treated approximately as a constant. Neglecting the friction term, Eq. (9.129) is rewritten as

$$\frac{\ddot{\delta}_b}{3} + \frac{k^2 c_s^2}{R^2(t)}\left[\frac{\delta_b}{3} + (1+R_b)\Psi(t,\boldsymbol{k})\right] = 0. \tag{9.132}$$

The acoustic oscillation starts as soon as the baryon–photon fluid enters the horizon (which defines $t = 0$) and ends at decoupling ($t = t_{\text{dc}}$). $R_b(t)$ is given by Eq. (9.76) and grows from $R_b(0) \simeq 0$ (radiation-dominated) to $R_b(t_{\text{dc}}) \simeq 0.592$ (Eq. (9.79b)) at decoupling (matter-dominated). The photon density fluctuation is given by $\delta_\gamma = (4/3)\delta_b$, and the temperature anisotropy is given by $\Theta \equiv \delta T/T = \delta_b/3$.

---

[13] The relativistic equations for the baryon–photon fluid are given by [785–787]

$$\left[\frac{\partial^2}{\partial t^2} + \left(H + \frac{\dot{R}_b}{1+R_b}\right)\frac{\partial}{\partial t} + \frac{c_s^2 k^2}{R^2}\right](\Theta + \Phi) = \frac{c_s^2 k^2}{R^2}\left[\Phi - (1+R_b)\Psi\right] \tag{9.130}$$

$$\dot{\Theta} + \frac{i}{3}\frac{k}{R}v_{\gamma,k} = -\dot{\Phi}. \tag{9.131}$$

For a perfect liquid, the space metric satisfies $\Phi = -\Psi$. Under the assumption of matter dominance, $\dot{\Psi}$ is constant. This can be seen from Eq. (9.125). The density grows as $\rho_0 \sim R^{-3}$, $\delta \sim R$, and hence $\Psi \sim R^0 =$ constant. For a tightly coupled photon–baryon liquid, we can set $\Theta = \delta T/T = \delta_b/3$ and $v_\gamma = v_b$. Then (9.131) reduces to Eq. (9.25a). If one further assumes constant R_b and neglects the friction term, Eq. (9.130) reduces to Eq. (9.132).

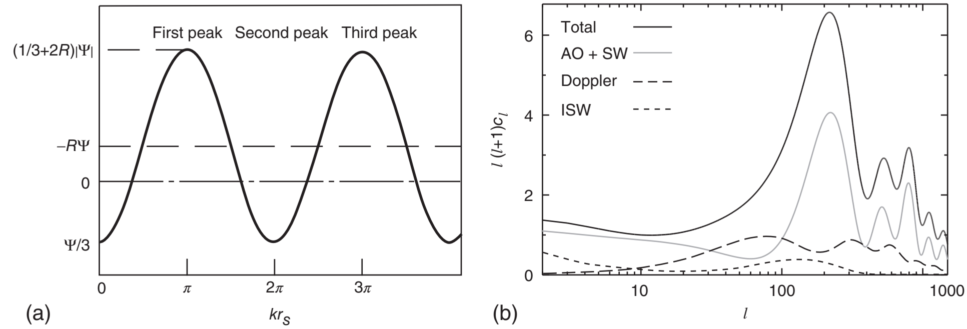


Figure 9.13 (a) Schematic drawing of the acoustic oscillation without damping. The oscillation center is offset at $-R_b\Psi$, and hence the first and third (compression) peaks are higher than the second (expansion) peak. (b) Contribution of the various terms to the temperature-anisotropy power spectrum from adiabatic initial conditions: (Total: thick line, AO (acoustic oscillation) + SW(Sachs–Wolfe); grey line; Doppler effect; dashed line and the integrated Sachs–Wolfe effect (ISW; short dahsed line). The units of the spectrum are arbitrary. (Reproduced with permission of [788].)

A general solution for Eq. (9.132) can be written as

$$\Theta = \frac{\delta_b}{3} = -(1 + R_b)\Psi + A\cos kr_s + B\sin kr_s \tag{9.133}$$

where $r_s = c_s t/R(t)$ is the sound horizon at time t. In the limit $k \to 0$, the temperature anisotropy at the decoupling is given by Eq. (9.123). As the potential is kept constant during the matter-dominated era (see Eq. (9.125)), it also provides the initial condition. We have $\Theta(0) = -(2/3)\Psi$. This condition determines $A = (1/3 + R_b)\Psi$. It is based on the assumption of adiabatic fluctuation which means that $\Theta = \delta_b/3 = \delta_\gamma/4$ is maintained through the super-horizon era to become the initial condition. In other words, $A \neq 0$, $\dot{\Theta}(0) = B = 0$.[14] As the observed anisotropy is sum of temperature fluctuation and gravitational potential, we have

$$\Theta + \Psi = \left(\frac{1}{3} + R_b\right)\Psi\cos kr_s - R_b\Psi \tag{9.134}$$

Oscillations inside the horizon start at the same time (coherent oscillation) and sum of those going in opposite directions behave as given by Eq. (9.134). Those modes having the right wavenumber to satisfy $kr_s = m\pi,\ m = 1, 2, \cdots$ make peaks. The mth peak with $m =$ odd corresponds to the maximum contraction and that with even m to maximum expansion (Figure 9.13a).

---

14) The opposite is the isocurvature or isothermal fluctuation, in which case $\delta\rho \propto \delta T/T = 0$ and $\Theta(0) = A = 0$, $\dot{\Theta} = B \neq 0$ in Eq. (9.133). In general, the solution is their mixture. A mixture of the isocurvature fluctuation changes the phase and hence the position of the acoustic peaks. Therefore, its contribution can be determined by measurements.

9.2.4
Doppler Effect

The Doppler effect is produced by the peculiar velocity of the source and the motion of the observer. Two kinds of velocities appear in the CMB.

$$\frac{\delta T}{T} = -\boldsymbol{v} \cdot \hat{\boldsymbol{k}}\Big|_{\text{source}} + \boldsymbol{v} \cdot \hat{\boldsymbol{k}}\Big|_{\text{observer}} \tag{9.135}$$

where $\hat{\boldsymbol{k}}$ is a unit vector in the direction of observation. It is taken positive for the direction to the source. The Doppler effect due to Earth's motion was already mentioned. It is observed as a dipole component of the CMB (see Figure 9.10b, i). This is the biggest component of the CMB ($\Delta T/T{\sim}10^{-3}$), and is about 100 times more intense than the other components. In the analysis, the dipole components are subtracted. The rest of the spectrum is very uniform, but anisotropy at the level of 10^{-5} is observed (see Figure 9.10b, ii and Figure 9.11). The peculiar velocity of the source can be approximately estimated as follows: For the tightly coupled photon–baryon liquid, $v_\gamma = v_b$ holds [786]. Then under the same assumption as we made to derive the acoustic oscillation, the nonrelativistic continuity equation for the baryon can be converted to that of the photon. With replacement of $\delta_b/3 \to \Theta$, $v_b \to v_\gamma$, Eq. (9.25a) becomes (see also footnote[13])

$$\dot{\delta} + \frac{\nabla \cdot \boldsymbol{v}}{R} = 0, \quad \to \quad \dot{\Theta} + \frac{1}{3}\frac{\nabla \cdot \boldsymbol{v}_{\text{photon}}}{R} = 0. \tag{9.136}$$

In the Fourier space, the velocity is expressed as

$$v_k = 3i\frac{R}{k}\dot{\Theta}. \tag{9.137}$$

Substituting Eq. (9.134) in Eq. (9.137), and using $kr_s = kc_s t/R$, we have

$$v_k = -ic_s\left(1 + 3R_b\right)\Psi \sin kr_s. \tag{9.138}$$

As the power spectrum is proportional to the square of the amplitude, it gives a subdominant contribution to the dominant acoustic oscillation term. As its phase is $\pi/2$ shifted relative to the dominant term, it tends to reduce the height of acoustic peaks (see Figure 9.13b).

9.2.5
Silk Damping

If the decoupling occurs instantly, their oscillation pattern will be imprinted as the initial condition for the matter density fluctuation growth and appears as resonance patterns above the power spectrum of the density fluctuation. In reality, the photons collide less frequently as the number of neutral hydrogens increases. Consequently, they diffuse as their mean free path becomes longer. Fluctuations with smaller scales than the diffusion length are reduced, which, in turn, act as a damping force to the baryon oscillation. Therefore, as the photon decouples, the oscillation is damped at scales smaller than the photon diffusion length. This is referred to as the *Silk damping* [789]. The diffusion length can be determined as the distance that the photon's random walk can reach within the Hubble time. The random walk is due to multiple Thomson scattering with electrons. Its mean free path is given by

$$l_{\text{mean}} = \frac{1}{n_e \sigma_T} \tag{9.139}$$

where σ_T is the total Thomson scattering cross section, and n_e is the number density of the electron. As the scattering is isotropic and the random walk is completely statistical, the reach of a random walk due to N collisions is given by $\sqrt{N} l_{\text{mean}}$ where N is given by H^{-1}/l_{mean}. Consequently, the diffusion length is given by

$$\begin{aligned} l_{\text{diff}} &= \sqrt{N}\, l_{\text{mean}} = \sqrt{\frac{H^{-1}}{l_{\text{mean}}} l_{\text{mean}}} = \left(\frac{3}{8\pi G_N \rho_0}\right)^{\frac{1}{4}} (l_{\text{mean}})^{\frac{1}{2}} \\ &= \left(\frac{3}{8\pi G_N \rho_0}\right)^{\frac{1}{4}} \left(\frac{m_N}{\rho_b \sigma_T}\right)^{\frac{1}{2}} = \left(\frac{3}{8\pi G_N}\right)^{\frac{1}{4}} \left(\frac{m_N}{\sigma_T}\right)^{\frac{1}{2}} \frac{1}{\rho_0^{\frac{1}{4}} \rho_b^{\frac{1}{2}}} \end{aligned} \tag{9.140}$$

where $n_b = \rho_b/m_N$ was used in deriving the penultimate equality. As the universe is already matter-dominated,

$$\rho_0 \simeq \rho_m = (1+z_{\text{dc}})^3 \rho_{m0} = (1+z_{\text{dc}})^3 \Omega_{m0} \rho_c, \qquad \rho_b = (1+z_{\text{dc}})^3 \Omega_{b0} \rho_c. \tag{9.141}$$

We have

$$l_{\text{diff}} = \left(\frac{3}{8\pi G_N}\right)^{\frac{1}{4}} \left(\frac{m_N}{\sigma_T}\right)^{\frac{1}{2}} \frac{(1+z_{\text{dc}})^{-\frac{9}{4}}}{\Omega_{m0}^{\frac{1}{4}} \Omega_{b0}^{\frac{1}{2}} \rho_c^{\frac{3}{4}}} = \left(\frac{8\pi G_N}{3} \frac{m_N}{\sigma_T}\right)^{\frac{1}{2}} \frac{(1+z_{\text{dc}})^{-\frac{9}{4}}}{H_0^{\frac{3}{2}} \Omega_{m0}^{\frac{1}{4}} \Omega_{b0}^{\frac{1}{2}}} \tag{9.142a}$$

It is referred to as the *Silk damping length*. In the comoving coordinate frame, it is given by

$$L_{\text{diff}} = l_{\text{diff}}(1+z_{\text{dc}}) = 19 \left(\frac{1+z_{\text{dc}}}{1091}\right)^{-\frac{5}{4}} \left(\frac{\Omega_{b0}}{0.045}\right)^{-\frac{1}{2}} \left(\frac{\Omega_{m0}}{0.27}\right)^{-\frac{1}{4}} \text{Mpc.}^{15)} \tag{9.142b}$$

Accordingly, the anisotropy with scale $\lambda < L_{diff}$ ($l \gtrsim 1000$) is damped as is observed in Figure 9.12 (see also Figure 9.14). The mass contained in this diffusion scale (Silk mass) is given by

$$M_{\text{Silk}} = \frac{4}{3}\pi \left(\frac{L_{\text{diff}}}{2}\right)^3 \rho_{b0} \simeq 2.2 \times 10^{13}\, M_\odot \tag{9.144}$$

---

15) A more precise phenomenological formula for the Silk scale k_{Silk} is given by [759]

$$\begin{aligned} k_{\text{Silk}} &= 1.6\,(\Omega_{b0} h^2)^{0.52} (\Omega_{m0} h^2)^{0.73} \times [1 + (10.4\,\Omega_{m0} h^2)^{-0.95}] \text{Mpc}^{-1} \\ k_{\text{Silk}}^{-1} &= 8.38\,\text{h}^{-1}\text{Mpc} \;\; [790] \end{aligned} \tag{9.143}$$

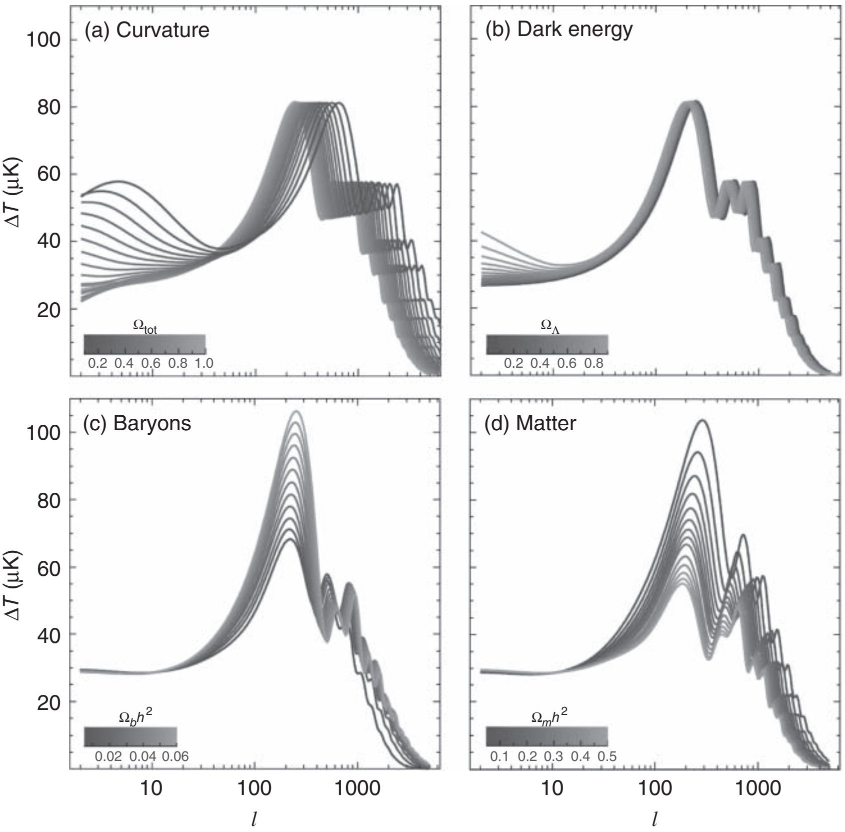


Figure 9.14 Sensitivity of the acoustic temperature spectrum to four fundamental cosmological parameters: (a) The curvature as quantified by Ω_{tot}. (b) The dark energy as quantified by the cosmological constant $\Omega_\Lambda (w = -1)$. (c) The physical baryon density $\Omega_{b0} h^2$. (d) The physical matter density $\Omega_{m0} h^2$, all varied around a fiducial model of $\Omega_{\text{tot}} = 1$, $\Omega_\lambda = 0.65$, $\Omega_{b0} h^2 = 0.02$, $\Omega_{m0} h^2 = 0.147$,n $= 1, z_{ri} = 0, E_i = 0$. (Reproduced with permission of [791].) (Please find a color version of this figure on the color plates.)

which exceeds the galaxy mass. That is, the density fluctuations with scale smaller than the galaxy scale were erased at the time of recombination.

9.2.6
Outcome of CMB Measurements

Curvature of the Universe The first peak is positioned at the sound horizon at the time of decoupling. It was estimated in Eq. (9.80). Here we give a simple derivation. Its size is approximately given by

$$D_s = R(t) r_s = R(t_{\text{dc}}) \int_0^{t_{\text{dc}}} dt \, \frac{c_s}{R} = R(t_{dc}) \int_0^{R_{\text{dc}}} \frac{c_s dR}{R^2 H(R)} = \frac{c_s R_0}{1 + z_{\text{dc}}} \int_{z_{\text{dc}}}^{\infty} \frac{d\tilde{z}}{H(\tilde{z})}. \quad (9.145)$$

For approximate evaluation, we put $c_s \sim 1/\sqrt{3}$ and $H(z) \simeq H_0\sqrt{\Omega_{m0}}(1+z)^{\frac{3}{2}}$ (see (8.68)), and obtain

$$D_s \sim \frac{2}{\sqrt{3}} \frac{R_0}{H_0 \Omega_{m0}^{\frac{1}{2}}(1+z_{\rm dc})^{\frac{3}{2}}}. \tag{9.146}$$

The angular diameter distance to the surface of last scattering is given by Eq. (8.74)

$$\begin{aligned} d_A &= \frac{R_0}{1+z_{\rm dc}} S_k(\chi_{\rm dc}) \overset{\text{for } k=0}{=\!=} \frac{R_0}{1+z_{\rm dc}} \chi_{\rm dc} \\ \chi_{\rm dc} &= \int_0^{r_{\rm dc}} \frac{dr}{\sqrt{1-kr^2}} = \int_0^{t_{\rm dc}} \frac{dt}{R(t)} \overset{\Omega_{m0}(1+z)\gg 1}{\simeq} \frac{2}{R_0 H_0 \Omega_{m0}^{\frac{1}{2}}}. \end{aligned} \tag{9.147}$$

The last approximation means that the distance from the Big Bang to the surface of the last scattering is negligible compared to the distance to the present epoch. Then, the angle of vision to the acoustic horizon for a flat matter-dominated universe is given by

$$\theta \simeq \frac{D}{d_A} \simeq \frac{1/\sqrt{3}}{\sqrt{1+z_{\rm dc}}} \simeq 0.0174 \simeq 1^\circ. \tag{9.148}$$

It corresponds to $l \simeq \pi/\theta \sim 200$, and reproduces the position of the first peak approximately. As is shown by Eq. (9.147), the angular diameter distance changes as a function of the cosmic curvature. That is, one can determine the curvature of the universe from the first peak position (see Figure 9.14a). Another indicator is the late ISW effect at the super-horizon scale ($l \lesssim 30$).

Dark Energy Composition $S_k(\chi)$ in Eq. (9.147) depends on the relative contribution of the dark energy (i.e., Ω_Λ). Figure 9.14b shows the dark energy dependence of the acoustic peaks. The peak position does not move much. Its behavior at low l is similar to the curvature effect due to late ISW effect. But as we noted earlier, the cosmic variance is large at small l. Hence, the dark energy cannot be determined very well from the CMB spectrum. More precise determination of the dark energy composition can be deduced from observation of accelerating cosmic expansion which is discussed in Section 11.1.

Baryon Density from CMB As the oscillation center is shifted by $-R_b\Psi$, amplitudes at odd peaks and even peaks are different. As Eq. (9.134) shows, strength of the first peak is $\sim|-(1/3+2R_b)\Psi|^2$ but that of the second peak is $\sim|(1/3)\Psi|^2$. The ratio of the first peak to the second is larger for larger baryon density. Qualitatively speaking, the larger baryon density corresponds to a smaller acoustic velocity. This is because the restoring force due to the pressure is weaker and the amplitude grows (see Figure 9.14(c)). The effect is to reduce the acoustic horizon and shifts the position of the first peak. The larger baryon density also reduces the diffusion length (see Eq. (9.144)). The result of CMB analysis on the baryon-to-photon ratio was already given in Figure 8.7.

Dark Matter Composition If the amount of the dark matter increases, the matter–radiation equality occurs earlier, which makes the matter component at decoupling larger. It reduces the sound speed, which, in turn, reduces the first peak. The increase of the matter component enhances the values of the Hubble parameter, which reduces the angle of view (see Figure 9.14(d)). That is, the existence of the dark matter changes both the peak position and its size, and hence the acoustic spectrum can determine its value accurately.

Reionization The universe today is highly ionized. As the photon recombination left all hydrogens and heliums neutral, it is of interest to know when they are reionized, that is, when the dark age ends. When the first stars are formed, they emit energetic photons and reionize neutral hydrogens (and also helium); in other words, we are interested to know when the first stars are formed. Quasars are found at distances up to $z < 7$ and their spectra of Lyman α series show no appreciable signs of absorption lines that would have been there if the neutral hydrogens existed on their line of path to the Earth. Reionization liberates free electrons and the CMB photons are scattered by them (Thomson scattering). Defining the optical depth $\tau_{ri} \equiv \int c dt\, n_e \sigma_{\text{Thomson}}$, the number of initial photons will be reduced to $N_\gamma = N_\gamma(0)e^{-\tau_{ri}}$, while the number of scattered photons will increase to $N_\gamma(0)(1 - e^{-\tau_{ri}})$. The scattered photons wash out the anisotropy, which we have parametrized as Θ. Therefore, the temperature distribution we observe would be modified to

$$T(1 + \Theta)e^{-\tau} + T(1 - e^{-\tau}) = T(1 + \Theta e^{-\tau}). \tag{9.149}$$

The reduction by reionization is effective only for those small-scale photons that are in the horizon at the time of reionization. Those at the super-horizon (roughly $l \lesssim 100$) are not affected. By investigating the CMB spectrum deformation, WMAP/Planck reported a value of $\tau_{ri} \simeq 0.084 \pm 0.016$, which corresponds to reionization at $z \simeq 10.9 \pm 1.4$ [123, 705]. Notice, however, that there are correlations among parameters. Deformations due to tilt (Eq. (9.100)) and reionization are very similar. What is usually done is to make an overall fit of the parameters or see changes of a parameter of interest by fixing others. Quasars are observed up to a distance of $z = 7$, and their number drastically reduces beyond $z > 7$. The reionization is likely to have taken place sometime $7 < z \lesssim 11$.

Other Cosmological Parameters Using the ΛCDM model, the WMAP and subsequently the Planck determined many important cosmic parameters very accurately [123, 674, 705]. Some of them are given in Table 9.1.

The WMAP/Planck observations made no new discovery, but the accuracy and overall consistency of the data were so overwhelming that they elevated the ΛCDM model with parameters some of which are listed in Table 9.1 (referred to as the *concordance model*) to the status of the SM of cosmology. It is an epoch making event analogous to establishment of the SM in particle physics in the 1970s.

9.2.7
Polarization

The CMBs are electromagnetic waves and can have polarizations. Generally, polarization is generated when the wave is reflected or when passing through a birefringent material. For the CMB, it is generated by Thomson scattering. It is caused by anisotropy of the density fluctuation or, more accurately, by its quadrupole component. It is also generated by gravitational waves. The measurement of the polarization provides valuable information in many ways. First, it can detect the imprint of the primordial gravitational waves produced by the inflation. That is, it provides a means to probe the formation of the universe before the Big Bang. Second, as polarization can be effected only by the last scattering, it can separate the anisotropy generated at the surface of the last scattering from that by reionization which happened much later. The latter occurs at much smaller z, leaving the imprint at much larger scale. Third, it can provide independent information on cosmological parameters and resolve degeneracies that cannot be separated from analysis of the temperature anisotropy alone.

Thomson Scattering Thomson scattering has differential cross section expressed as

$$\frac{d\sigma}{d\Omega} = \frac{\sigma_T}{8\pi}(\boldsymbol{\epsilon}_f \cdot \boldsymbol{\epsilon}_i). \tag{9.150}$$

Here, $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_f$ are the polarizations of the photon before and after scattering. This has the property that the initial polarization in the direction of scattering is completely erased.

Figure 9.15 illustrates how the polarization ε_f after scattering from the electron is related to the polarization ε_i before scattering. If the density distribution of the photon is uniform, polarizations induced by photons coming from above and

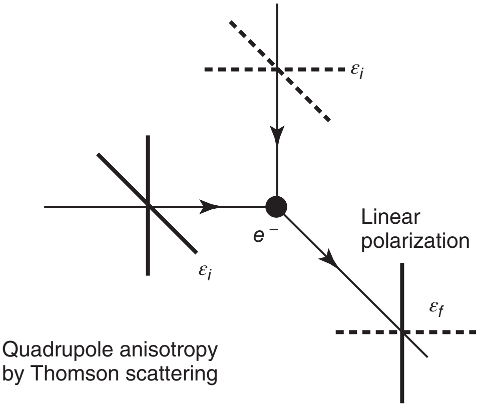


Figure 9.15 Initial polarization in the direction of scattering is completely erased. The illustration shows that only vertical polarization from the photon entering sideways and only horizontal polarization from the photon entering downward survive.

those from side are equal, and no net polarization is produced after scattering. But if the density has a quadrupole moment component, the scattered photons are polarized. If the scatterings are repeated, no net polarizations remain. Therefore, for the polarizations to be produced, the number density of electrons must be large enough to scatter photons but small enough to allow the scattered photons to escape from additional scattering. Photons at the time of recombination have mean free paths long enough to escape from the scatterer.

Therefore, only those from the last scattering are polarized. This means that polarizations appear at the same scale as the Silk damping, that is, at $l \sim 100 - 1000$, and are damped at larger or smaller scales. Only about 10% of the photons are polarized because of the diminishing density at the decoupling. Therefore, polarization measurements are not easy experimentally. After decoupling, the photon has chances to be polarized again by reionized electrons at $z \lesssim 11$. As the distance is much closer to us, this polarization is made at larger scale (peaking around $l \sim 10$). Therefore, it is easy to separate the effect of reionization. In fact, WMAP/Planck used the scale position of the polarization to determine the time of reionization. Additional polarizations can be induced by gravitational lensing in the path.

E Mode and B Mode When a monochromatic wave propagates in the $\boldsymbol{k}$ direction, the electric field vector is expressed as

$$\mathbf{E}(t, \boldsymbol{x}) = (\mathbf{e}_x E_x + \mathbf{e}_y E_y) e^{i\boldsymbol{k}\cdot\boldsymbol{x} - i\omega t} \tag{9.151}$$

where $\mathbf{e}_x$, $\mathbf{e}_y$ are the unit vectors in the x, y direction. Generally E_x, E_y are complex and can be expressed as $E_x = |E_x| e^{i\delta_x}$, $E_y = |E_y| e^{i\delta_y}$. If $\delta_x = \delta_y$, the field is linearly polarized. Its amplitude is given by $E = \sqrt{E_x^2 + E_y^2}$ and its azimuthal angle by $\theta = \tan^{-1}(E_y/E_x)$. If $\delta_x = 0$, $\delta_y = \pi/2$, it has elliptical polarization with major and minor axes on the x, y direction; otherwise, it represents an elliptical polarization with its axes rotated. The circular polarization is obtained for $|E_x| = |E_y|$ and $\delta_y - \delta_x = \pi/2$. In general, the polarization can be represented by a polarization tensor [792, 793]

$$\boldsymbol{T} = [t_{ij}] = [<(\mathbf{e}_i^* \cdot \mathbf{E})(\mathbf{e}_j \cdot \mathbf{E})>] \equiv \frac{1}{2}\begin{pmatrix} I+Q & U-iV \\ U+iV & I-Q \end{pmatrix} \tag{9.152}$$

$$I \equiv <|E_x|^2 + |E_y|^2> \tag{9.153a}$$

$$Q \equiv <|E_x|^2 - |E_y|^2> \tag{9.153b}$$

$$U \equiv 2 <\mathrm{Re}|E_x^*||E_y|> \cos(\delta_y - \delta_x) \tag{9.153c}$$

$$V \equiv 2 <\mathrm{Im}|E_x^*||E_y|> \sin(\delta_y - \delta_x) \tag{9.153d}$$

where $<\cdots>$ means the statistical ensemble of the measured data. I, Q, U, V are referred to as the *Stokes parameters*. $I = \mathrm{Tr}[\mathbf{T}]$ is the intensity of the electromagnetic field and is proportional to the fourth power of the temperature T^4 for the CMB. If one describes polarizations using "rods" on a map, the magnitude is given by $P = \sqrt{Q^2 + U^2}$ and the orientation makes an angle $\alpha = (1/2)\tan^{-1}(U/Q)$.

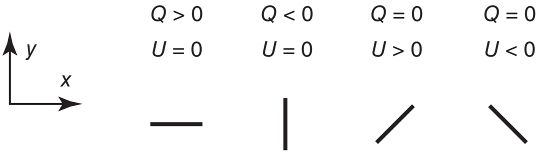


Figure 9.16 Relations between Q, U, and orientation of the polarization.

V represents circular polarization. As the Thomson scattering can only produce linear polarizations, we may ignore it for the following discussions. Relations between orientations of the polarization and the parameters Q, U are illustrated in Figure 9.16.

By rotation in a plane perpendicular to the direction $\hat{\boldsymbol{n}} = \boldsymbol{k}/|\boldsymbol{k}|$ by an angle ψ, that is, by changing the unit vectors to $\mathbf{e}'_x = \cos\psi\mathbf{e}_x + \sin\psi\mathbf{e}_y$, $\mathbf{e}'_y = -\sin\psi\mathbf{e}_x + \cos\psi\mathbf{e}_y$, I dose not change but Q and U rotate. Every time the coordinate system is rotated by 45°, the variables change as $Q \to U \to -Q \to -U \to Q$. Such observables dependent on the coordinate system are inconvenient to analyze. We notice that under rotation by angle ψ, the Q and U transform as

$$(Q + iU)(\hat{\boldsymbol{n}}) \to (Q + iU)'(\hat{\boldsymbol{n}}) = e^{\pm 2i\psi}(Q + iU)(\hat{\boldsymbol{n}}). \tag{9.154}$$

The rotational property Eq. (9.154) for the Q and U means that the quantities $Q \pm iU$ have spin 2. As I is a scalar under rotation, it has spin 0. Statistical properties of the radiation field were analyzed after decomposing it in terms of spherical harmonics as was done in Eq. (9.110). However, it can be used only for scalar fields such as the temperature field. The combination $Q \pm iU$ are spin-2 objects and the same formula cannot be applied. For spin s objects, we can use generalized tensor spherical harmonics denoted as ${}_sY_{lm}$. In terms of them, the temperature anisotropy $\Theta = \delta T/T = (1/4)\delta I/I$ and Q, U are expanded as [788, 794] (see also Appendix F)

$$\Theta(\hat{\boldsymbol{n}}) = \sum_{l=2}\sum_{m} a^T_{lm} Y_{lm}(\theta, \phi), \tag{9.155a}$$

$$(Q \pm iU)(\hat{\boldsymbol{n}}) = \sum_{l=2}\sum_{m} (a_{\pm 2,\,lm})\,\{{}_{\pm 2}Y_{lm}(\theta, \phi)\}. \tag{9.155b}$$

The functions ${}_{\pm 2}Y_{lm}$ are related to the standard spherical harmonic functions by

$${}_{\pm 2}Y_{lm}(\theta, \phi) = -\left[\frac{(l-2)!}{(l+2)!}\right]\sin^2(\theta)\left[\frac{\partial}{\partial\theta} \pm i\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\right]\sin^{-2}\theta\; Y_{lm}(\theta, \phi). \tag{9.156}$$

Coefficients of the expansion can be determined using the completeness and orthonormality of the ${}_sY_{lm}$ functions.

$$\begin{aligned} a^T_{lm} &= \int d\Omega\; Y^*_{lm}(\theta, \phi)\Theta(\hat{\boldsymbol{n}}), \\ a_{\pm 2,\,lm} &= \int d\Omega\; {}_{\pm 2}Y^*_{lm}(\theta, \phi)(Q \pm iU)(\hat{\boldsymbol{n}}). \end{aligned} \tag{9.157}$$

They satisfy the relations $a^*_{-2,\,lm} = a_{2,\,lm}$ to ensure the reality of Q and U. Instead of $a_{\pm 2,\,lm}$, it is more convenient to introduce their linear combinations

$$a_{E,\,lm} = -\frac{1}{2}(a_{2,\,lm} + a_{-2,\,lm}), \qquad a_{B,\,lm} = -\frac{1}{2i}(a_{2,\,lm} - a_{-2,\,lm}). \tag{9.158}$$

$$E(\hat{\boldsymbol{n}}) \equiv \sum_{lm} a_{E,lm} Y_{lm}(\theta, \phi), \qquad B(\hat{\boldsymbol{n}}) \equiv \sum_{lm} a_{B,lm} Y_{lm}(\theta, \phi) \tag{9.159}$$

The advantage of using E and B instead of U and Q parameters is that in the Fourier space, the former is rotation-invariant. All the rotating parts are absorbed in the tensor harmonics. Thus, the $E(\hat{\boldsymbol{n}})$ and $B(\hat{\boldsymbol{n}})$ functions defined by Eq. (9.159) are scalars in real space [795–797]. Therefore, they can be treated exactly as the temperature anisotropy functions. Besides, they have definite parity. The E mode has even parity and the B mode has odd parity. This fact makes the scalar components of the quadrupole (i.e., those that are expressed in terms of $\nabla_i v_j \sim \nabla_i \nabla_j$) free of the B modes. To generate B-mode patterns, a tensor potential is needed. Notice that the gravitational waves are in tensor modes (see Section E.4). Thus, by using E and B functions, one can reject the main component of the temperature anisotropy which works as background in extracting the gravitational wave effects.

We illustrate how the polarizations parameterized by E and B functions appear for the case of plane waves in Figure 9.17 and radial waves in Figure 9.18. Let us digress a bit, and see why they show such patterns.

The trough of the wave represents cold spots and the crest the hot spots. The cold spot is where the density is maximum, and the hot spot is where the density is minimum. The direction of momentum flow is from the hot to the cold spots. This is counterintuitive. But the observed anisotropy is the sum of temperature fluctuation and gravitational potential. If the gravitational potential overwhelms the

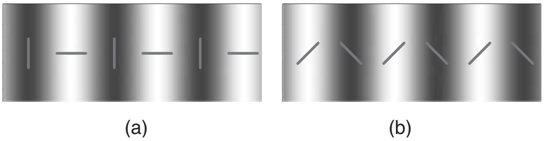


Figure 9.17 Patterns of E- and B-mode polarizations for plane waves. Polarization patterns for an E mode (a) and B mode (b) on a small patch of the sky for potentials that are locally Fourier modes. The shading denotes the amplitude of the potential and the headless vectors denote the unsigned direction of the polarization. For the electric pattern, the polarization is aligned along or perpendicular to the Fourier wave vector depending on the sign of the potential. For the magnetic pattern, the polarization is at 45 deg.

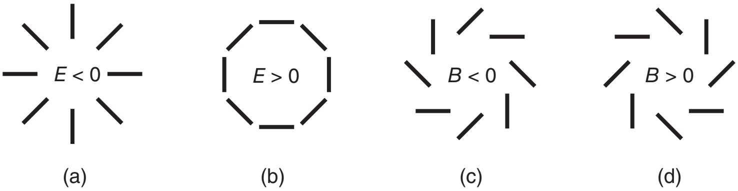


Figure 9.18 Polarization fields for radial waves. Positive (negative) E-modes surround hot (cold) spots. B-modes cannot be produced by ordinary perturbations to the density but are produced by gravity waves. Note that E (B) polarizations are rotational invariants but even (odd) under reflection.

pressure, which is the case for the CMB for super-horizon scale, the flow is toward the potential minimum, that is, from hot spots to cold spots. For scales below the horizon, sometimes pressure overwhelms the gravitational force. In such a case, the hot spot is made by the pressure and momentum flow is directed to the hot spots. For simplicity, we consider the former case.

Suppose an observer is in the middle of the trough; light coming from upwards or downwards and scattered toward the observer is more intense than that coming from the sides because it comes from the hot spots. Therefore, the polarization is horizontal or parallel to the momentum flow in this case. If the observer is at the top of the crest, reverse is true and the polarization is vertical or transverse to the flow, as illustrated in Figure 9.17(a).

The E-mode polarization cannot produce slant polarization like those in Figure 9.17(b). This is because of the even parity nature of the E mode. While both E and B mode polarizations posses rotational symmetry on the xy plane, the E mode is even with respect to x or y reflection, which allows vertical or horizontal polarization but forbids slant polarization. The reverse is true for the B-mode polarization. Polarization patterns for radial waves are illustrated in Figure 9.18.

The polarization field has spin 2, hence it comes back to the original position when rotated by 180 deg while the ordinary vector needs 360 deg to come back. Nevertheless, the E and B polarizations share many common features with the electric field E and magnetic field B. In the context of anisotropies induced by the density fluctuations, the quadrupole is generated by gradients of the velocity. The E polarization is produced by components of the velocity gradient $\sim\partial_i v_i$, while the B polarization is produced by the rotational mode of the velocity $(\sim\partial_i v_j)$. The important point is that polarizations produced by the density fluctuation (scalar source) are constrained to be the E mode. This is because the velocity produced by the density anisotropy is longitudinal, that is, $v_i \propto k_i \sim \nabla_i$. Then the quadrupole component is proportional to $\partial_i\partial_j$, which is symmetric in ij and no rotational component is produced.

The gravitational waves can produce both E- and B-mode polarizations. Therefore, detection of the B mode provides a clean separation of the primordial gravitational waves.

The power spectra are defined as rotationally invariant quantities. If the fluctuations are Gaussian, the power spectra are independent of magnetic quantum number m. Then, to characterize the statistics of the CMB perturbations, only four power spectra are needed.

$$\begin{aligned}
C_l^{\mathrm{TT}} &= \frac{1}{2l+1}\sum_m < a^*_{\mathrm{T},lm} a_{\mathrm{T},lm} > \\
C_l^{\mathrm{EE}} &= \frac{1}{2l+1}\sum_m < a^*_{\mathrm{E},lm} a_{\mathrm{E},lm} > \\
C_l^{\mathrm{BB}} &= \frac{1}{2l+1}\sum_m < a^*_{\mathrm{B},lm} a_{\mathrm{B},lm} > \\
C_l^{\mathrm{TE}} &= \frac{1}{2l+1}\sum_m < a^*_{\mathrm{T},lm} a_{\mathrm{E},lm} >
\end{aligned} \tag{9.160}$$

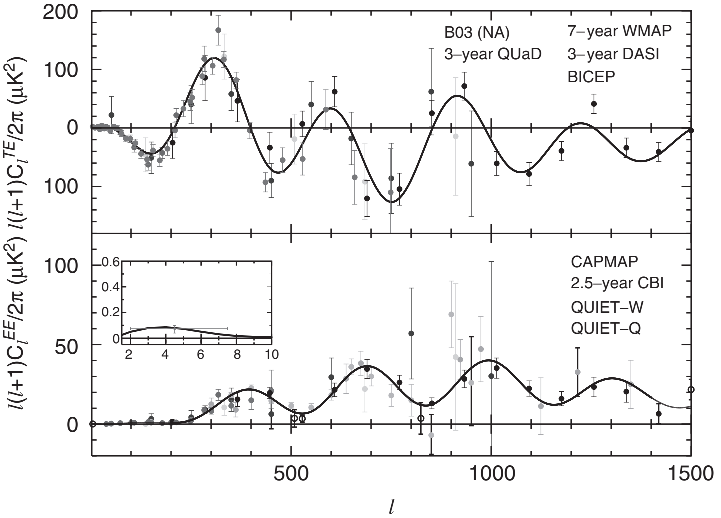


Figure 9.19 Current measurements of the polarization power spectra TE (top) and EE. [704, 799–805] (Reproduced with permission of [798].)

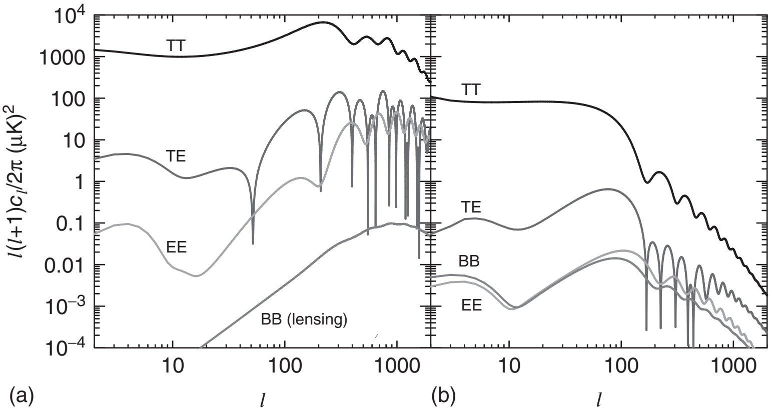


Figure 9.20 (a) CMB power spectra from scalar perturbations of temperature, E mode, B mode, and T-E cross-correlation. The B-mode spectrum induced by weak gravitational lensing is also shown. (b) Tensor perturbations (gravitational waves). The amplitude of the primordial gravitational wave at the maximum amplitude allowed by current data (the tensor-to-scalar ratio $r = 0.22$ [705]). The bump at $l \sim 5$ is due to the reionization effect. (Reproduced with permission of [798].)

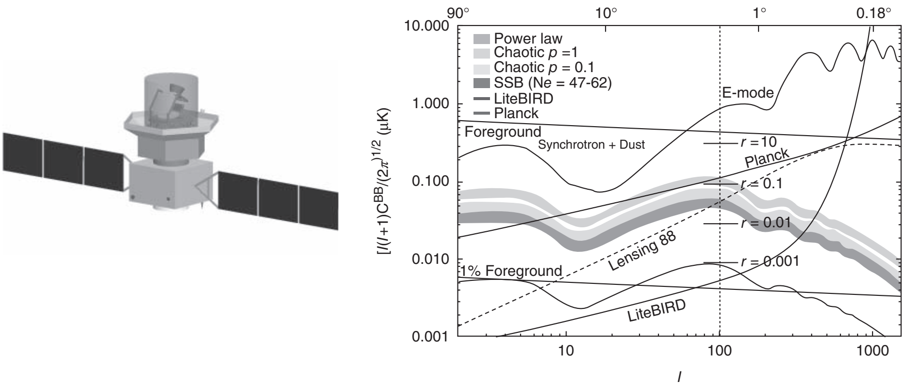


Figure 9.21 LiteBIRD image and its reach of the B-mode polarization in CMB. Dashed line indicates B mode by lensing. The anticipated sensitivity is ×100 that of the Planck, however effective suppression of the foreground is essential for its success. The four bands correspond to models listed on upper left. (Reproduced with permission of [809].)

The cross-correlations between B and E or B and T vanish if there are no parity-violating interactions.

Status of B-mode Measurements The polarization that has its origin in CMB density fluctuations exhibits the same oscillating structure as the temperature anisotropy. While the density fluctuations is somewhat obscured by the out-of-phase Doppler effect, the polarization is free of such effect and has sharper peaks. Only TE and EE modes have been observed so far (Figure 9.19), and were shown to be consistent with the SM.

The BB modes are still challenges for the experimentalists. The amplitude for the gravitational waves is directly proportional to the square of the energy scale of the inflation. Their strength is about 10 times the slow rolling parameter ϵ (see Eq. (9.104)) but it is highly model-dependent (see discussions in Section 9.1.9). Theoretical predictions assume some value of the tensor to scalar ratio $r = P_t(k_0)/P_s(k_0)$, where $P(k_0)$ is the power spectrum at a certain wavenumber ($k_0 = 0.002\ \mathrm{Mpc}^{-1}$ [705]).

Figure 9.20(b) shows an example of predictions on gravitational wave effects compared to those in scalar modes shown in Figure 9.20(a) [798]. A value of $r = 0.22$ was assumed in calculating the tensor spectra.

The expected sensitivities of future projects expressed as the tensor-to-scalar ratio are about $r{\sim}0.022$ for PolarBear (ground-based [806]), $r \simeq 0.01$ for CMBPOL (CMB polarization satellite [807]), and $r = 0.001$ for LiteBIRD (Lite satellite for the studies of B-mode polarization and Inflation from cosmic background Radiation Detection) [808]. Figure 9.21 illustrates the sensitivity of LiteBIRD which was proposed to JAXA (Japan Aerospace Exploration Agency) and is expected to start in ~2020.

10
Dark Matter

In the previous two chapters, we saw that the advent of cosmology clarified the existence of two unknown entities, namely the dark matter (DM) and the dark energy (DE). The newly established Standard Model (SM) of cosmology says that the known matter (the SM particles) occupies only about 5% of the total cosmic energy budget, the dark matter occupies about 26%, and the rest is occupied by the dark energy. They are new challenges for particle physicists. They are undoubtedly the main themes of particle physics in the twenty-first century. In this chapter, we discuss their properties, their candidates, and some means to detect them.

10.1
Cosmic Budget

The dark matter was discovered as something that is not accounted for by the known matter. By matter, we mean baryons most of the time. We know that stars are made of baryons. Knowing the matter composition of the universe starts with counting how many stars are there. But as can be seen from Table 10.1, amount of matter in the stars occupy only 15% of the total baryons. Most of them exist in the form of interstellar medium (ISM), which are made of gas (ionic, atomic, and molecular), dust, and cosmic rays. Hot plasma is a source of X-rays. Warm plasma emits electromagnetic radiation of longer wavelengths, but most of them are cold and do not emit light. Adding up all the baryons, the fraction of the baryon density in the universe ($\Omega_b = \rho_{b0}/\rho_{c0}$) amounts to about 2% of the total density. When compared with the baryon density $\Omega_{b0} = 0.0487 \pm 0.0006$ as obtained from the Big Bang Cosmology [123], it reproduces the correct value within a factor 2.

10.2
Evidences of Dark Matter

Historically, the existence of the dark matter was first suggested by Zwicky [811] in 1933 to explain the anomalously violent motion of galaxies in the Coma cluster. However, it had been forgotten for a long time until it was revived in the 1970s by observations of rotational velocity of spiral galaxies using the 21-cm bright lines of hydrogen [812]. Today, many clusters of galaxies are known to contain the dark

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

Table 10.1 Baryon budget in the universe in units of $\Omega = \rho/\rho_c$ [810].

Component	Central (observed value at $z \approx 0$)	Maximum	Minimum	Grade
1. Stars in spheroids	$0.0026\ h_{70}^{-1}$	$0.0043\ h_{70}^{-1}$	$0.0014\ h_{70}^{-1}$	A
2. Stars in disks	$0.00086\ h_{70}^{-1}$	$0.00129\ h_{70}^{-1}$	$0.00051\ h_{70}^{-1}$	A^-
3. Stars in irregulars	$0.000069\ h_{70}^{-1}$	$0.000116\ h_{70}^{-1}$	$0.000033\ h_{70}^{-1}$	B
4. Neutral atomic gas	$0.00033\ h_{70}^{-1}$	$0.00041\ h_{70}^{-1}$	$0.00025\ h_{70}^{-1}$	A
5. Molecular gas	$0.00030\ h_{70}^{-1}$	$0.00037h_{70}^{-1}$	$0.00023h_{70}^{-1}$	A^-
6. Plasma in clusters	$0.0026\ h_{70}^{-1.5}$	$0.0044\ h_{70}^{-1.5}$	$0.0014\ h_{70}^{-1.5}$	A
7a. Warm plasma in groups[a]	$0.0056\ h_{70}^{-1.5}$	$0.0115\ h_{70}^{-1.5}$	$0.0029\ h_{70}^{-1.5}$	B
7b. Cool plasma	$0.002\ h_{70}^{-1}$	$0.003\ h_{70}^{-1}$	$0.0007\ h_{70}^{-1}$	C
7'. Plasma in groups[b]	$0.014\ h_{70}^{-1}$	$0.030\ h_{70}^{-1}$	$0.0072\ h_{70}^{-1}$	B
8. Sum ($h = 70$, $z \simeq 0$)	0.021	0.041	0.007	. . .

The Hubble constant is defined by $70\,\text{h}_{70}\text{km s}^{-1}\ \text{Mpc}^{-1}$.

[a] 7a, associated with groups of galaxies and 7b estimate by Lyα.

[b] In principle, $7' = 7a + 7b$, $7'$ is adopted for sum.

matter. The total mass of a typical cluster is 10–100 times that of visible galaxies in it [813]. Other evidences are given by X-ray sources broadly distributed within clusters of galaxies. They are emitted by thermal motion of particles, and to confine them locally by gravity, accumulation of 10–100 times more mass of visible stars is necessary. Recently, the advent of the gravitational lensing technique has made it possible to measure the mass distribution directly without resorting to visual observations. Theoretically, the dark matter is needed at least in two cases. One is to stabilize disks in simulations for forming spiral galaxies [814, 815], and the other is to match the evolution speed of cosmic structure formation (see Eq. (9.48) and discussions afterward). And of course, there is the precise determination of the baryon density Ω_b (Eq. (8.132)) using the Hot Big Bang ΛCDM model, which turned out to be much less than the total matter density Ω_{m0}. We give short explanations about some of them below.

10.2.1
Rotation Curves of Spiral Galaxies

Observations of characteristic 21-cm radio waves that hydrogen atoms emit revealed that the rotation velocities at the periphery of spiral galaxies stay constant far beyond boundaries of visible disks. The behavior of the rotational velocities can be predicted from Newton's law of energy conservation under gravity:

$$\frac{1}{2}mv^2 - G_N\frac{mM}{r} = E \lesssim 0 \quad \rightarrow \quad v = \sqrt{\frac{2G_N M}{r}}. \tag{10.1}$$

One can estimate the mass M of the spiral galaxy by counting the number of luminous stars in it, and expect the velocity to decrease $\sim\sqrt{1/r}$ beyond the boundary

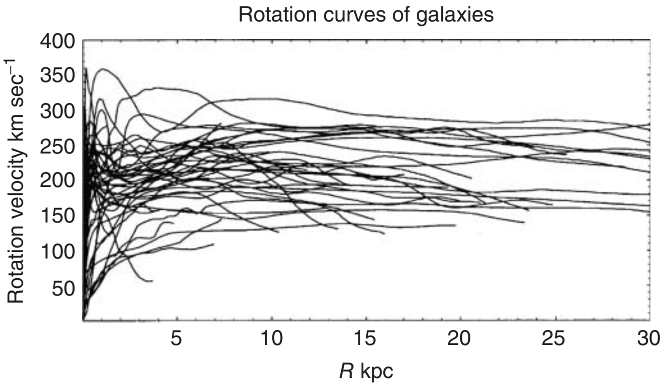


Figure 10.1 Rotation velocities of spiral galaxies obtained by combining CO data for the central regions, optical for disks, and HI (an interstellar cloud composed of neutral atomic hydrogen) for outer disk and halo. The graph shows that the constancy of the velocity goes beyond 30 kpc, while the typical radius of disks in spiral galaxies is ~10 kpc. (Reproduced with permission of [818, 819]).

of the disks. Typically, the disk's radius is of the order ~10 kpc. Observations show that, in many spiral galaxies, the rotation velocity $v \simeq$ constant up to 30–40 kpc with no sign of decrease yet [816, 817]. Figure 10.1 shows a compilation of rotation curves of a large sample of galaxies (see also Figure 10.14).

This means that the mass attracting the rotating objects increases with r. It is estimated that the whole mass of the spiral galaxies is at least 10 times that of visible stars. Elliptical galaxies also show evidences for the dark matter via strong gravitational lensing and/or emission of extended X-rays.

10.2.2
Virial Mass of the Clusters

Clusters of galaxies are the largest astronomical objects that are in bound states. Their mass can be estimated by several methods, including application of the virial theorem, study of X-ray emission profile, and weak gravitational lensing.

Here we start describing the virial mass. Motions of member galaxies in a cluster are considered as random and almost thermal. Whether the galaxies are in complete thermal equilibrium is debatable, but we assume that they are random enough to make the following arguments plausible. If particles are bound by gravity and in thermal equilibrium, the virial theorem says that (see (A.9))

$$
\begin{aligned}
&<E> = <K> + <U>, \quad <K> = -\frac{1}{2} <U> \\
&<K> = \frac{m}{2} <v^2> = \frac{G_N}{2} m \int \frac{dM}{r}
\end{aligned}
\tag{10.2}
$$

Where $<E>$, $<K>$ and $<U>$ are the average total, kinetic and potential energy of each particle.

Consider a cluster whose total mass M is distributed spherically and isotropically. Assuming that the system is in thermal equilibrium, its constituent galaxy has average energy given by

$$< K > = \frac{1}{2} < mv^2 >= -\frac{1}{2} < U >= \frac{3}{10} G \frac{mM}{R}, \tag{10.3a}$$

$$< v^2 > = < v_r^2 > + < v_\theta^2 > + < v_\phi^2 >= 3 < v_r^2 > \tag{10.3b}$$

$$\therefore \quad M = M_{\text{virial}} \equiv \frac{5R < v_r^2 >}{G}. \tag{10.3c}$$

$< v_r^2 >$ is the velocity dispersion in the direction of observation and is a measurable quantity. Consequently, by measuring the mean velocity dispersion $< v_r^2 >$ and the average distance R, the total mass of the cluster can be deduced. The mass defined this way is referred to as the *virial mass*. Observations show that, typically, the virial mass of the clusters is up to ~100 times the visible mass.

10.2.3
X-ray Emitting Clusters

Very often, clusters of galaxies are also X-ray emitters (Figure 10.2). X-rays are emitted by Bremsstrahlung of energetic ($< K > \sim$ 10 KeV) baryons in hot plasma with $T \sim 100 \times 10^6$ K. To confine such fast moving particles, a large gravitational potential is necessary. From the intensity or total energy of the X-rays, one can calculate the total mass to bind the hot plasma gas. It amounts to $1 \sim 5$ times the visible mass in the X-ray emitting region. The total mass of the clusters can be calculated to be $\sim 10^{14\sim15}\ M_\odot$. Contribution of the visible stars amounts to ~2%. The mass of the clusters derived using X-rays nearly agrees with the virial mass.

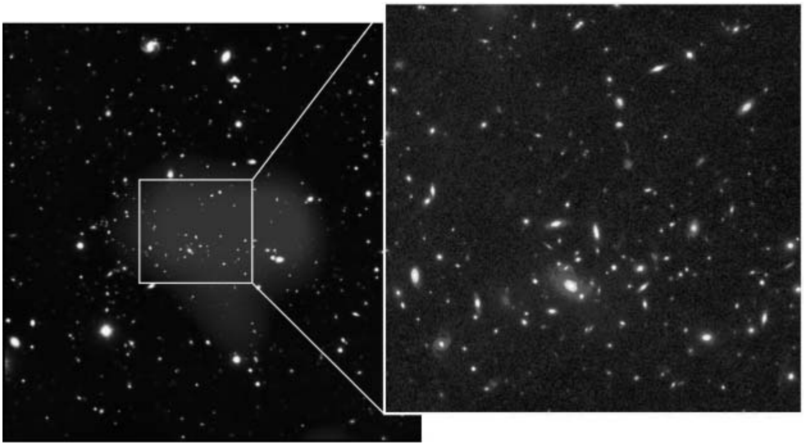


Figure 10.2 Cluster of galaxies MS1054 – 0321 at ~2.5 Gpc distance ($z = 0.831$). Its diameter is ~3 Mpc, and it contains thousands of galaxies. Left: X-ray sources (blue colored) are overlayed with the cluster pictures in the visible range. Right: Enlarged image by the Hubble telescope. (Reproduced with permission of [820–822].) (Please find a color version of this figure on the color plates.)

Problem 10.1 *Assume that X-rays are emitted uniformly from a spherically distributed cluster with $R = 3$ Mpc. Assuming that the mean kinetic energy of baryons that emit X-rays to be* ~ 1 *keV and using Eq.* (10.3c)*, we can derive the mass of the cluster. Show that it amounts to* $M \sim 6.6 \times 10^{14} M_\odot$*. If one assumes the galaxy mass as* $2 \times 10^{11} M_\odot$*, about 3300 galaxies are contained in the cluster.*

10.2.4
Gravitational Lens

Fermat's Principle Gravitational lensing has recently become one of the principal probes of the dark matter and the dark energy [823–826].

It responds to genuine matter distribution and is not biased by the M/L (mass to luminosity) ratio, a nuisance which has to be dealt with if one relies on traditional means of weighing shining objects. If the method is applied to a local sample of lensing galaxies, it provides additional means of measuring the local mass distribution independent of optical or radio telescopes. Its advantage (especially of the weak lensing) comes from the fact that shears of galaxy images can be measured straightforwardly. Thus a large number of galaxies can be sampled easily. While galaxy shear measurements themselves do not provide dark energy information, correlation of galaxy shear across the sky does, especially if red-shift information of galaxies is utilized to obtain 3-D "tomography." This is because the existence of the dark energy affects the structure formation process and thus distribution of lenses provides valuable information. One byproduct of lensing is local time delay of light so that fluctuations in the distant light sources appear at different times, sometimes hundreds of days apart on the Earth. It offers an alternative method to determine the Hubble constant. In order to derive laws of optical deflection in the framework of general relativity, we use the metric defined in Eq. (9.10). The light signal satisfies $ds^2 = 0$ hence its path length is given by

$$cdt = \sqrt{\frac{1-2\Psi}{1+2\Psi}}R(t)dr \simeq (1-2\Psi)ds \tag{10.4}$$

Comparing the above equation with that of flat space, we obtain the effective index of refraction.

$$n(\boldsymbol{x}) = \frac{c}{v_{\text{eff}}} = \frac{c}{\frac{ds}{dt}} = 1 - 2\Psi = 1 + |2\Psi|. \tag{10.5}$$

The Fermat's principle in optics tells that the light path minimizes $\int n\,dl$, that is, the deflection angle γ is given by the integral of the gradient of n perpendicular to the direction of propagation along the light path:

$$\gamma = -\int \nabla_\perp n dl = 2\int \nabla_\perp \Psi dl. \tag{10.6}$$

If the light is deflected by a mass M with a transverse distance ξ from the light path, the potential is given by (see Figure 10.3)

$$\Psi = -\frac{G_N M}{\sqrt{\xi + z^2}} \tag{10.7}$$

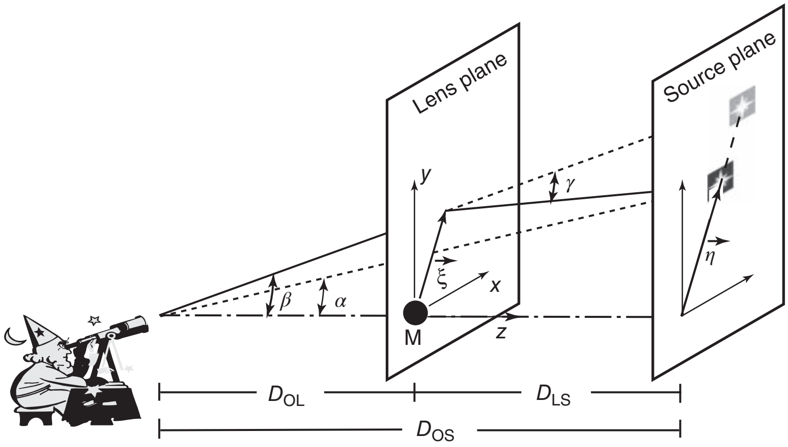


Figure 10.3 Geometry of gravitational lensing. The direction of the source in the absence of lens (M) is α. The observed direction is β.

where z is the longitudinal distance measured from the mass center position of the lens. Substituting Eq. (10.7) into Eq. (10.6), we obtain the total deflection angle.

$$\gamma = -2 \int_{-\infty}^{+\infty} \nabla_{\rho} \frac{G_N M}{\sqrt{\rho^2 + z^2}} dz \bigg|_{\rho=\xi} = \frac{4G_N M}{\xi}^{\,1)} \tag{10.8}$$

which is twice that given by Newtonian mechanics. For the Sun which has the Schwarzschild radius 2.95 km and the radius 6.96×10^5 km, a light ray grazing the limb of the Sun is deflected by 1, 7'' (''=arc sec). Eddington's confirmation of the effect in 1919 was a historic event that placed Einstein in the limelight.

Referring to Figure 10.3, one can derive the true angle of view α from the observed angle β. First, we notice that

$$\alpha D_{\text{OS}} = \beta D_{\text{OS}} - \gamma D_{\text{LS}} \tag{10.9}$$

where D_{OS}, D_{OL}, D_{LS} are the distances from the observer to the source and the lens, and between the lens and the source, respectively. Substituting Eq. (10.8) into Eq. (10.9) and using $\xi = \beta D_{\text{OL}}$, we obtain the lens equation

$$\alpha = \beta - \frac{D_{\text{LS}}}{D_{\text{OS}} D_{\text{OL}}} \frac{4G_N M}{\beta} \tag{10.10}$$

Note that the lens equation was obtained assuming Euclidean geometry and a static universe. For the expanding universe with curvature, the longitudinal lengths $D_{\text{OS}}, \ldots$ are to be replaced by the angular distances defined by Eq. (8.74). In general, the lens equation may have more than one solution, that is, the source object may have multiple images. For the special case of $\alpha = 0$, the azimuthal symmetry

---

1) As the Schwarzschild radius is given by $R_S = 2G_N M/c^2$, a good way to memorize the deflection angle is twice the inverse of the impact parameter in units of R_S.

requires that the image is ring-like with angular radius

$$\beta_E = \sqrt{4G_N M \frac{D_{\text{LS}}}{D_{\text{OS}} D_{\text{OL}}}}. \tag{10.11}$$

This is referred to as the *Einstein ring*. For $\alpha \neq 0$, one obtains two solutions for a point mass lens, that is

$$\beta_{\pm} = \frac{1}{2}\left(\alpha \pm \sqrt{\alpha^2 + 4\beta_E^2}\right) \tag{10.12}$$

with $\beta_+\beta_- = -\beta_E^2$. The minus sign means that the two images are on opposite sides of the lensing mass. If the source is circular, the two elongated arcs appear diametrically opposite to the center of the lens.

To get some feeling about the magnitude of lensing, we list a few numbers:

- For a star of 1 $M_\odot$ at distance $D = 10$ kpc, $\beta_E \approx 0.001''$; relevant for MACHO observation (see Section 10.3.4);
- For a cluster of $10^{14} M_\odot$ at $z = 0.3$ and sources at $z_S = 1$, $\beta_E \approx 20''$.

Strong Lensing and Weak Lensing If the angle α is comparable to β_E, we talk about strong lensing. In this case, double images of the arcs or the Einstein rings are visible. An example of strong lensing is given in Figure 10.4.

On the other hand, if $\alpha \gg \beta_E$, so that the two solutions for β Eq. (10.12) become

$$\beta_+ \to \alpha \gg \beta_E, \qquad \beta_- \to \frac{-\beta_E^2}{\alpha} \ll \beta_E, \tag{10.13}$$

then we talk of weak lensing. The β_- solution becomes invisible, while the β_+ solution nears an image almost without gravitational deflection and brightness magnification. Images of distant galaxies reach us only slightly deformed, but millions of samples are easily collected providing us the statistical means to

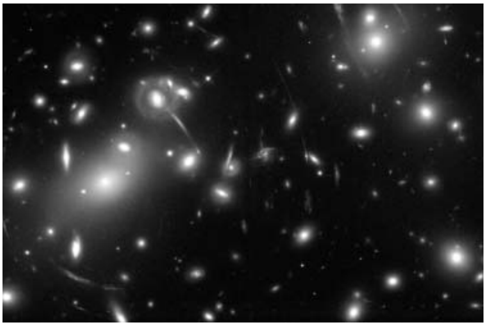


Figure 10.4 The cluster Abell 2218 hosts one of the most impressive collections of arcs. This HST (Hubble space telescope) image of the cluster's central region shows a pattern of strongly distorted galaxy images tangentially aligned with respect to the cluster center, which lies close to the bright galaxy in the left part of this image. $z = 0.176$ (~2.5 G lyrs) away in the northern constellation of the Dragon (Draco). Picture by NASA (Reproduced with permission of [827].)

process them. It provides a very promising application. Study of distorted images of distant galaxies created by inhomogeneous mass distribution in the light path will give us information on the dark matter distribution and hence the structure and evolution of the universe. Here we concentrate on the aspect of the weak lensing as a probe of the dark matter.

Let us consider now an extended source distributed around a point η on the source plane (see Figure 10.3). If we move the source point around η, the point where the light intersects the lens plane also moves around $\vec{\xi}$. In this case, the lens equation becomes a vector mapping ($\boldsymbol{\beta} \rightarrow \boldsymbol{\alpha}$) equation

$$\boldsymbol{\alpha} = \boldsymbol{\beta} - 4G_N M \frac{D_{\mathrm{LS}}}{D_{\mathrm{OS}} D_{\mathrm{OL}}} \frac{\boldsymbol{\beta}}{|\beta|^2}. \tag{10.14}$$

If the source is much smaller than the angular scale at which the lens properties change appreciably, the lens mapping can be linearized. It is useful to introduce a transfer matrix by which the image position shift $\Delta\boldsymbol{\beta} = (\Delta\beta_x, \Delta\beta_y)$ can be mapped to that of the source position $\Delta\boldsymbol{\alpha} = (\Delta\alpha_x, \Delta\alpha_y)$

$$\Delta\alpha_i \, (i = 1, 2) = \sum_{j=1}^{2} \frac{\partial\alpha_i}{\partial\beta_j} \Delta\beta_j \equiv \sum_j A_{ij} \Delta\beta_j \tag{10.15}$$

where the transfer matrix is given by

$$\begin{aligned} A_{ij} &= \delta_{ij} - 4G_N M \frac{D_{\mathrm{LS}}}{D_{\mathrm{OS}} D_{\mathrm{OL}}} \left(\frac{\delta_{ij}}{|\boldsymbol{\beta}|^2} - \frac{2\beta_i\beta_j}{|\boldsymbol{\beta}|^4} \right) \\ &= \delta_{ij} - 4G_N M \frac{D_{\mathrm{LS}} D_{\mathrm{OL}}}{D_{\mathrm{OS}}} \left(\frac{\delta_{ij}}{|\boldsymbol{\xi}|^2} - \frac{2\xi_i\xi_j}{|\boldsymbol{\xi}|^4} \right) \equiv \delta_{ij} - B_{ij}. \end{aligned} \tag{10.16}$$

It is conventional to express the reduced transfer matrix B_{ij} (referred to as the *shear matrix*) as

$$\boldsymbol{B} \equiv \begin{bmatrix} \kappa + \gamma_1 & \gamma_2 \\ \gamma_2 & \kappa - \gamma_1 \end{bmatrix} \tag{10.17}$$

with κ referred to as the *convergence* and γ_i known as the *shear field*. It will soon be shown that the convergence contains information on the mass concentration. As Eq. (10.16) shows, the shear matrix is expressed in terms of the Newtonian potential.

A more general formula of Eq. (10.14) and (10.16) for a distributed lenses based on the geodesic equation of general relativity is given by [787]

$$\alpha_i = \beta_i + 2 \int_0^{\chi_S} d\chi_L \, \frac{\partial\Psi[\boldsymbol{x}(\chi_L)]}{\partial\xi_i} \left(1 - \frac{\chi_L}{\chi_S} \right), \tag{10.18a}$$

$$B_{ij} = 2 \int_0^{\chi_S} d\chi_L \frac{\partial^2\Psi[\boldsymbol{x}(\chi_L)]}{\partial\xi_i\partial\xi_j} \chi_L \left(1 - \frac{\chi_L}{\chi_S} \right) \tag{10.18b}$$

where $\chi = \int dt/R(t)$ is the comoving distance from an observer ($\chi = 0$), and χ_L, χ_S are the comoving distances to the lens and the source. The position of the lens is given by $\boldsymbol{x}(\chi_L) = \boldsymbol{\xi} + \hat{\boldsymbol{n}}\chi_L$, where $\boldsymbol{\xi} = (\xi_1, \xi_2)$ is the two-dimensional vector and $\hat{\boldsymbol{n}}$ is the unit vector in the direction orthogonal to the (ξ_1, ξ_2) plane. It is related to the red shift "z" via the relation $d\chi = dz/H(z)$. We show that Eqs. (10.18) reproduces Eq. (10.14) and (10.16).

Proof: For simplicity, we consider a flat ($k = 0$) universe. Then, $\chi = r$ is the radial coordinate in the comoving frame. The gravitational potential on the light path (ξ_1, ξ_2, r) emitted by a point mass at $(0, 0, r_L)$ is given by

$$\Psi(\xi_1, \xi_2, r) = -\frac{G_N M}{R_L\left[(r - r_L)^2 + |\boldsymbol{\xi}|^2\right]^{\frac{1}{2}}} \tag{10.19}$$

where $R_L = R(t_L)$ is the scale factor at the lens plane. Substituting Eq. (10.19) in Eq. (10.18a), we have

$$\alpha_i = \beta_i - \int_0^{r_S} dr \left(1 - \frac{r}{r_S}\right) \frac{2G_N M \xi_i}{R_L\left[(r - r_L)^2 + |\boldsymbol{\xi}|^2\right]^{\frac{3}{2}}}. \tag{10.20}$$

For the cosmological scale, the function $\left[(r - r_L)^2 + |\boldsymbol{\xi}|^2\right]^{-3/2}$ is sharply peaked at $r = r_L$. Then the slowly varying function, that is, $(1 - r/r_S)$, can be replaced with its value at $r = r_L$ and can be taken out of the integral. The integration is easily carried out and gives

$$\alpha_i = \beta_i - 4G_N M \frac{(r_S - r_L)}{r_S r_L R_L} \frac{\beta_i}{|\boldsymbol{\beta}|^2} \tag{10.21}$$

where $\xi_i = r_L \beta_i$ is used in converting $\xi \to \beta_i$. As the angular distances are given by (see Eq. (8.74))

$$D_{\mathrm{LS}} = R(t_S)(r_S - r_L), \quad D_{\mathrm{OS}} = R(t_S) r_S, \quad D_{\mathrm{OL}} = R(t_L) r_L = R_L r_L \tag{10.22}$$

Equation (10.21) agrees with Eq. (10.14). Derivation of Eq. (10.16) can be performed similarly. ■

In many cases, it is not practical to measure the red shifts of individual galaxies whose images are distorted by lensing. In this case, the averaging must also include sampling along the line. For this purpose, we introduce a probability function $W(\chi_S)$ for the distribution of source galaxies along the line of sight. It is normalized by

$$\int_0^\infty d\chi_S W(\chi_S) = 1. \tag{10.23}$$

Then the transfer matrix is modified to take into account the effective averaging along the line of sight, which leads to

$$\begin{aligned} B_{ij}(\hat{n}) &= \int_0^\infty d\chi_S\, W(\chi_S) B_{ij}(\chi_S, \hat{n}) \\ &= 2\int_0^\infty d\chi_S W(\chi_S) \int_0^{\chi_S} d\chi_L \left(1 - \frac{\chi_L}{\chi_S}\right) \chi_L \left(\frac{\partial^2}{\partial\xi_i \partial\xi_j}\Psi\left[\boldsymbol{x}(\chi_L)\right]\right) \qquad (10.24) \\ &= \int_0^\infty d\chi_L \left(\frac{\partial^2}{\partial\xi_i \partial\xi_j}\Psi\left[\boldsymbol{x}(\chi_L)\right]\right) g(\chi_L) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{x}(\chi) &= \boldsymbol{\xi} + \chi\hat{\boldsymbol{n}} = (\xi_1, \xi_2, \chi) \\ g(\chi) &= 2\chi \int_\chi^\infty d\chi_S \left(1 - \frac{\chi}{\chi_S}\right) W(\chi_S). \end{aligned} \tag{10.25}$$

For a flat universe, $\chi_L(1 - \chi_L/\chi_S)$ in Eq. (10.24) can be replaced with $r_L(1 - r_L/r_S)$.

Now, let us consider the extraction of the convergence κ. It can be obtained as the trace of the matrix **B** (see Eq. (10.17)). Using Eq. (10.18b) and $\chi = r$ for the flat universe, the convergence becomes

$$\kappa = \frac{1}{2}\mathrm{Tr}B = \int_0^{r_S} dr\, \frac{r(r_S - r)}{r_S} \left(\nabla^2 - \frac{\partial^2}{\partial r^2}\right) \Psi(\boldsymbol{\xi}, r). \tag{10.26}$$

We consider a collection of mass, that is, a cluster of galaxies which are located more or less at the same distance at $r \simeq r_L$. As the potential falls off rapidly for large $|r - r_L|$, the slowly varying function $r(r_S - r)$ can be replaced with $r_L(r_S - r_L)$ and taken out of the integral. Then the second term in parentheses gives a vanishing contribution. The first term can be re-expressed in terms of the mass density fluctuation using the Poisson equation $\nabla^2\Psi = 4\pi G_N R_L^2 \delta\rho_M$. Thus

$$\kappa(\boldsymbol{\xi}) = 4\pi G_N \frac{D_{\mathrm{LS}} D_{\mathrm{OL}}}{D_{\mathrm{OS}}} \int dr\, R_L \delta\rho_M(\boldsymbol{\xi}, r). \tag{10.27}$$

The integral is taken for a range of r passing through the lens. Thus the convergence κ gives information of the total mass density of a cluster of lensing galaxies that lie along the line of sight.

At this point, before discussing how to measure κ, it is relevant to show how the knowledge of κ was used to detect the existence of the dark matter.

Figure 10.5a gives an example of measuring the mass distribution using the weak gravitational lensing method. Compare this with the X-ray image of the same spot in Figure 10.5b. Note that contours of the gravitational potential coincide with the location of visible galaxies, but not with the location of the X-ray gas (produced by the hot plasma of baryons). The "bullet" is a smaller subcluster that has passed through the larger cluster. The hot gas created by the two colliding clusters are seen in X-rays, being clearly distinct from the two mass clusters, giving a convincing evidence of the dark matter.

Shear Matrix We have learned that the convergence κ contains information on the mass concentration. It remains to show how to measure the convergence, namely the shear matrix. The shear matrix gives information about perturbations to the gravitational potential by mass distribution along the light path. We expect lensing to convert circular images to elliptical ones. To describe this effect, then, we need to quantify the measure of the ellipticity. We start from defining the quadrupole

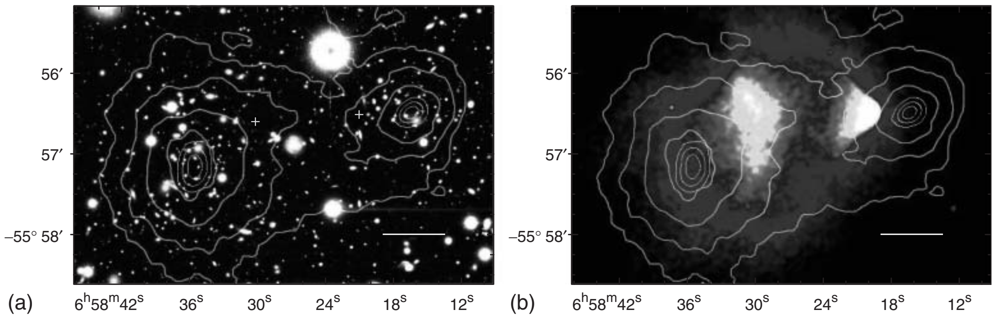


Figure 10.5 (a) Color image of the merging "bullet" cluster 1E 0657-558 obtained with the 6.5-m Magellan telescope in the Las Campanas Observatory, with the white bar indicating 200 kpc at the distance of the cluster. (b) Chandra X-ray image of the cluster. Shown in green contours in both panels are the weak-lensing κ reconstructions, with the outer contour levels at $\kappa = 0.16$ and increasing in steps of 0.07. The white contours show the errors at 68.3%, 95.5%, and 99.7% confidence levels. The two figures are interpreted as follows. The two clusters passed through each other 150 million years ago. The dark matter clump, revealed by the weak-lensing map, is coincident with the collisionless galaxies and lies ahead of the collisional gas. The gas of the smaller "bullet" cluster (right), creating a "bow-shaped shock wave located near the right side of the cluster, lags behind the cluster galaxies." (Reproduced with permission of [828].) (Please find a color version of this figure on the color plates.)

moments of an image.

$$Q_{ij} = \frac{\int d^2\beta I_{\mathrm{obs}}(\beta)\beta_i\beta_j}{\int d^2\beta I_{\mathrm{obs}}(\beta)} \tag{10.28}$$

A circular image has $Q_{xx} = Q_{yy}$ and $Q_{xy} = 0$. Therefore, two good measures of ellipticity are

$$\varepsilon_1 = \frac{Q_{xx} - Q_{yy}}{Q_{xx} + Q_{yy}}, \qquad \varepsilon_2 = \frac{2Q_{xy}}{Q_{xx} + Q_{yy}}. \tag{10.29}$$

To grasp physical image of the ellipticity, Figure 10.6 shows different patterns depending on the orientation of the ellipses.

Note also that two ellipses can align parallel to each other but cannot align themselves perpendicular to each other. For instance, the pattern in Figure 10.7a can appear when overdensity exists at the center and (b) appears when underdensity

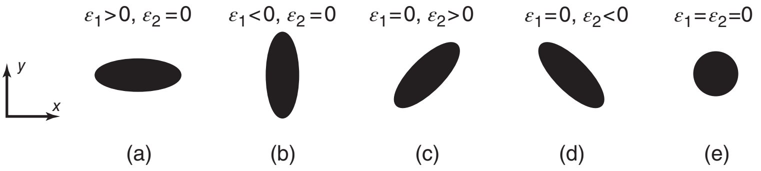


Figure 10.6 Shapes and orientation of ovals depending on values of ellipticity ε_1 and ε_2.

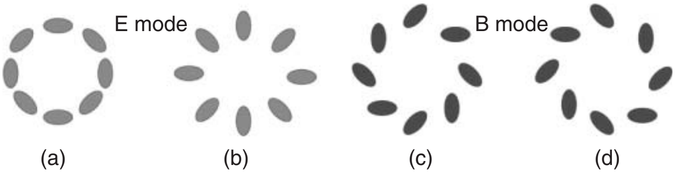


Figure 10.7 (a) E>0,(b) E<0 indicate shear patterns in rotation-free modes (E modes) allowed by gravitational lensing. (c) B<0,(d) B>0 indicate rotation modes (B modes) not allowed from a scalar gravitational potential. Only the E modes give the signal of the aperture mass statistics.

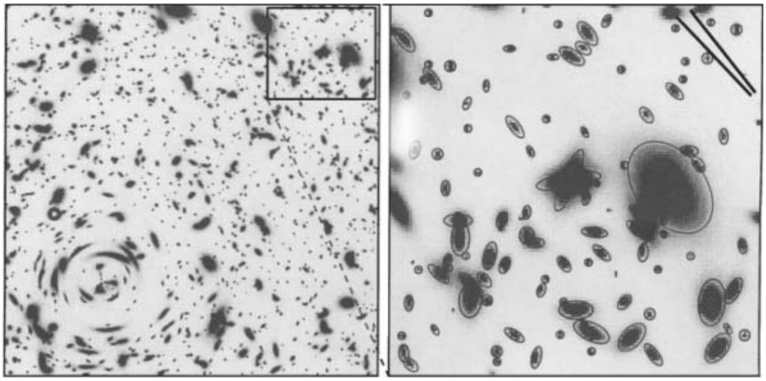


Figure 10.8 Illustration of the two lensing regimes. The left panel is a simulation of a cluster of galaxies at red shift 0.15, modeled by an isothermal sphere with a velocity dispersion of 1300 *km/s*. The lensed population has an average red shift of 1. In the innermost region (bottom left part of the panel), tangential and radial arcs are clearly identified. As the radial distance of lensed galaxies increases, the shear decreases, and far from the cluster center the ellipticity produced by the shear is lower than the intrinsic ellipticity of the galaxies. The lensing signal must be averaged over a large number of galaxies in order to be measured accurately. The right panel is a zoomed version of the inset in the left panel. It shows the images of the galaxies in the weak lensing regime. The contours show their shape as determined from their second moments. The average orientation of these galaxies is given by the solid lines at the top right. The lower line is the true orientation of the shear produced by the cluster at that position, and the upper line is the orientation computed from 92 galaxies of the zoomed area. The difference between the two orientations is random noise due to the intrinsic ellipticity and orientation distributions of the galaxies. (Reproduced with permission of [823].)

exists. But patterns (c) and (d) are not allowed.[2] Both patterns (a) and (b) have $\varepsilon_1(\vec{\theta}_1)\varepsilon_1(\vec{\theta}_2) > 0$, but (c) and (d) are characterized by negative values $\varepsilon_2(\vec{\theta}_1)\varepsilon_2(\vec{\theta}_2) < 0$. This fact can be utilized to eliminate the background noise statistically.

Figure 10.8 shows a simulation of how correlations between shear orientations and mass distributions are related for strong and weak lenses [829].

---

[2] This is because the origin of the deformation is a scalar potential. To produce B-mode patterns, a tensor potential is needed. This fact was discussed in the detection of gravitational waves by observing the cosmic microwave background (CMB) photon polarizations [See Section 9.2.7].

How can one obtain the shear matrix from observations? As Eq. (10.28) shows, the quadrupole moments and the ellipticity ε_1, ε_2 are directly measurable quantities. We need formulas to extract shear information from observations. One distinct feature of the gravitational lens is that it is purely geometrical; it does not change the energy spectrum or the number of photons. Hence the surface brightness is kept constant. The total luminosity depends only on the geometrical area of the images. Thus we have

$$I_{\text{obs}}(\beta) = I_{\text{source}}(\alpha). \tag{10.30}$$

Then

$$\int d^2\beta I_{\text{obs}}(\beta)\,\beta_i\,\beta_j = \sum_{kl} (A^{-1})_{ik}(A^{-1})_{jl} \det A^{-1} \int d^2\alpha I_{\text{source}}(\alpha)\,\alpha_k\alpha_l, \tag{10.31a}$$

$$\int d^2\beta\, I_{\text{obs}}(\beta) = \det A^{-1} \int d^2\alpha I_{\text{source}}(\alpha) \tag{10.31b}$$

where the matrix A_{ij} was given by Eq. (10.16). Therefore

$$Q_{ij}^{\text{obs}} = \sum_{kl} (A^{-1})_{ik}(A^{-1})_{jl} Q_{kl}^{S} \tag{10.32}$$

where

$$Q_{kl}^{S} = \frac{\int d^2\alpha I_{\text{source}}(\alpha)\alpha_k\alpha_l}{\int d^2\alpha I_{\text{source}}(\alpha)} \tag{10.33}$$

is the quadrupole matrix in the absence of lensing. Incidentally, the left-hand side of Eq. (10.31b) is the total brightness of the image and the integral of the right-hand side is that of the source. Thus the magnification of the lens is given by

$$\mu = \frac{\int d^2\beta\, I_{\text{observed}}(\beta)}{\int d^2\alpha\, I_{\text{source}}(\alpha)} = \frac{1}{\det A} = \frac{1}{(1-\kappa)^2 - \gamma^2} \tag{10.34}$$

where the shear amplitude $\gamma = \sqrt{\gamma_1^2 + \gamma_2^2}$ expresses the amplitude of the anisotropic magnification. On the other hand, κ expresses the isotropic part of the magnification and is directly related to the projected mass density, as was shown in Eq. (10.27).

For a circular source or for the average of a large sample, one may set $< Q_{ij}^{S} >= \overline{Q}\delta_{ij}/2$ with $\overline{Q}$ a constant which does not depend on the (ij) indices. Then

$$< Q_{ij} > = \sum_{k} (A^{-1})_{ik}(A^{-1})_{jl} \frac{\overline{Q}}{2}\delta_{kl} = \frac{\overline{Q}}{2}[(A^{-1})^2]_{ij}. \tag{10.35}$$

We can eliminate the unknown $\overline{Q}$ by dividing with the trace of the averaged quadrupole moment. Referring to Eq. (10.35) and (10.32), we have

$$\varepsilon_1 = \frac{< Q_{11} > - < Q_{22} >}{\text{Tr}Q} = \frac{(A^{-1})_{11}^2 - (A^{-1})_{22}^2}{(A^{-1})_{11}^2 + (A^{-1})_{22}^2}. \tag{10.36}$$

Using

$$A^{-1} = \frac{1}{(1-\kappa)^2 - \gamma_1^2 - \gamma_2^2} \begin{bmatrix} 1-\kappa+\gamma_1 & \gamma_2 \\ \gamma_2 & 1-\kappa-\gamma_1 \end{bmatrix} \tag{10.37}$$

we obtain

$$\begin{aligned}\varepsilon_1 &= \frac{<Q_{11}> - <Q_{22}>}{\mathrm{Tr}Q} = \frac{2\gamma_1(1-\kappa)}{(1-\kappa)^2 + (\gamma_1^2 + \gamma_2^2)} \\ \varepsilon_2 &= \frac{2<Q_{12}>}{\mathrm{Tr}Q} = \frac{2\gamma_2(1-\kappa)}{(1-\kappa)^2 + (\gamma_1^2 + \gamma_2^2)}.^{3)}\end{aligned} \tag{10.38}$$

If all the distortions are small, then

$$\varepsilon_1 \simeq 2\gamma_1, \qquad \varepsilon_2 \simeq 2\gamma_2. \tag{10.40}$$

So far so good. But notice we can only extract the traceless part of the shear matrix. The value of κ that we are really interested in is the trace part which cannot be obtained by the above method. It can be obtained by looking at the gradients of the traceless part of the shear matrix [830]. The convergence and the shear fields are related through the transfer matrix by

$$\kappa = \frac{1}{2}\left(B_{11} + B_{22}\right), \qquad \gamma_1 = \frac{1}{2}\left(B_{11} - B_{22}\right), \qquad \gamma_2 = B_{12}. \tag{10.41}$$

Combining the above relations with Eq. (10.18b), we immediately obtain

$$\frac{\partial\kappa}{\partial\xi_1} = \frac{\partial\gamma_1}{\partial\xi_1} + \frac{\partial\gamma_2}{\partial\xi_2}, \qquad \frac{\partial\kappa}{\partial\xi_2} = -\frac{\partial\gamma_1}{\partial\xi_2} + \frac{\partial\gamma_2}{\partial\xi_1}. \tag{10.42}$$

Thus κ can be calculated from the measured shear field γ_i up to the direction-independent constant. Figure 10.9 shows an example of 3-D construction of clustering the dark matter [831, 832].

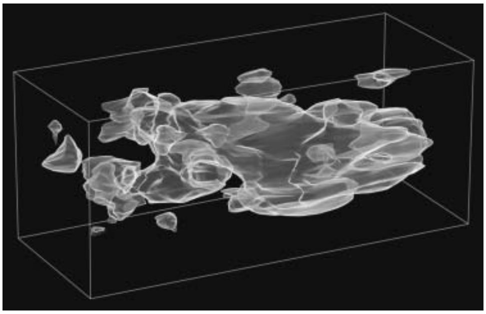


Figure 10.9 3-D image of the evolving distribution of the dark matter reconstructed using gravitational lensing. The distance from the Earth increases from left ($z \simeq 0$) to right ($z \simeq 1.1$) over a span of 25 Mpc. The clumping of the dark matter becomes more pronounced, moving right to left across the volume map, from the early universe to the more recent Universe. The distribution is calibrated by measuring the cosmological red shift ($0 < z < 1.1$) of the lensing galaxies in the constellation Sextans. The structure spans over more than 25 Mpc. Courtesy of NASA [833].

---

3) Some authors use "reduced complex shear g" instead of "shear matrix". Denoting $\gamma = \gamma_1 + i\gamma_2$ and $\delta = \varepsilon_1 + i\varepsilon_2$,

$$g = \frac{\gamma}{1-\kappa}; \quad \delta \equiv \varepsilon_1 + i\varepsilon_2 = \frac{2g}{1+|g|^2} = \frac{2\gamma(1-\kappa)}{(1-\kappa)^2 + |\gamma|^2}. \tag{10.39}$$

10.3
Relics of the Big Bang

Evidence of the dark matter is overwhelming. They are considered as relics of the Big Bang. They were produced in abundance when the cosmic temperature was much higher than the mass of the dark matter, but were left out of thermal equilibrium after their interaction rate could no longer catch up the Hubble expansion rate. Their number density at the decoupling is frozen and its energy spectrum is also frozen but is simply red-shifted as the universe expands. To detect them, we need to know their present number density, energy spectrum, and the interaction strength.

10.3.1
Freeze-Out

A rigorous treatment of the freeze-out has to be based on the Boltzmann transport equation in an expanding universe, but here we give a simplified treatment (see [125] for a more complete treatment). We consider a case of an unknown particle χ which can decay to X where X stands for any type of particle to which the χ's can annihilate. We consider a process

$$\chi + \overline{\chi} \to X + \overline{X} \tag{10.43}$$

In the thermal universe, if the number density n_χ is larger than its thermal equilibrium value n_χ^{eq}, more particles will be annihilated at the rate $n_\chi^2 v\sigma(\chi + \overline{\chi} \to X + \overline{X})$ and n_χ will decrease. In the expanding universe, the rate has to be modified from dn/dt to $[1/R(t)]^3[d(nR^3)/dt]$. Thus we obtain the basic equation that governs the departure from equilibrium for the species χ, as

$$\frac{dn_\chi}{dt} + 3Hn_\chi = -<v\sigma_\chi> [n_\chi^2 - (n_\chi^{eq})^2] \tag{10.44}$$

where $<\cdots>$ stands for thermal average which is necessary because the energy as well as the velocity direction at which the reaction occurs is random. The first term represents the reduction of n_χ by the reaction Eq. (10.43), and the second term the increase of n_χ by the inverse reaction. When there are more than one reaction channels, σ_χ is to be replaced with $\sum_i \sigma_{\chi,i}$. The suffix "eq" means that it is a number in thermal equilibrium. If we use the number of relic particles in a volume, $N_\chi = n_\chi V \propto n_\chi R^3$, Eq. (10.44) can be re-expressed in a form:

$$\frac{d\ln N_\chi}{d\ln R} = -\frac{\Gamma}{H}\left[1 - \left(\frac{N_\chi^{eq}}{N_\chi}\right)^2\right] \tag{10.45}$$

where $\Gamma = n_\chi <v\sigma_\chi>$ is the reaction rate. Equation (10.45) shows that the increase or decrease of the production is governed by the factor Γ_χ/H. As the total entropy in a volume $S = sV$ is a conserved number, the same equation holds if we divide N_χ by S, that is, replace N_χ by $Y_\chi = n_\chi/s = N_\chi/S$. This is the number density per entropy in a comoving volume and is a much more useful variable as will be shown shortly. Introducing a new variable $x = m_\chi/T$, we have a similar equation for Y:

$$\frac{d\ln Y_\chi}{d\ln x} = -\frac{\Gamma}{H}\left[1 - \left(\frac{Y_\chi^{eq}}{Y_\chi}\right)^2\right] \tag{10.46}$$

For $\Gamma \gg H$, we have a solution $Y_\chi \simeq Y_\chi^{eq}$. It is easy to understand its meaning. The reaction is fast and thermal equilibrium is maintained so that Y_χ follows closely that of thermal equilibrium Y_χ^{eq}. For $\Gamma \ll H$, the right-hand side can be neglected and we have $Y_\chi = Y_\chi(x \to \infty) \equiv Y_\chi^\infty = \text{constant}$. Here, the particle χ is decoupled from the reaction, hence Y_χ loses track of Y_χ^{eq} and $Y_\chi - Y_\chi^{eq} \gg Y_\chi^{eq}$.

For intermediate values of $\Gamma/H \sim O(1)$, the equation and the value of Y_χ^∞ have to be solved numerically. But it is known that a good approximation for Y_χ^∞ can be obtained by setting the freeze-out temperature $x = x_f$, where x_f can be determined by the condition $\Gamma = H$.

10.3.2
Hot Dark Matter

If the species χ is relativistic at freeze-out, say ($x_f = m_\chi/T_f < 3$), the relic is referred to as the *hot dark matter (HDM)*. Then Y_χ^{eq} does not change appreciably with time during the period of freeze-out because of the power dependence of the number density on the temperature with no Boltzmann reduction factor, and the resulting $Y_\chi(\infty)$ is just the equilibrium value at freeze-out. Using Eq. (8.39a) and (8.47b), we have

$$Y_\chi(\infty) = Y_\chi^{eq}(x_f) = \frac{n_\chi(T_f)}{s(T_f)} = \frac{g_{\text{eff}}\dfrac{\zeta(3)}{\pi^2}T_f^3}{g_s(T_f)\dfrac{2\pi^2}{45}T_f^3} = \frac{45\zeta(3)}{2\pi^4}\frac{g_{\text{eff}}}{g_s(T_f)} \tag{10.47}$$

where $g_{\text{eff}} = g_\chi$ for bosons and $3g_\chi/4$ for fermions. Sometimes, it is useful to express n_χ in terms of the photon number. Using $n_\gamma = [2\zeta(3)/\pi^2]T^3$ to re-express the entropy, the present number density of the hot relic particle can be expressed as

$$n_{\chi,\text{HDM}}^0 = s(T_0)Y_\chi^\infty = \frac{s(T_0)}{s(T_f)}n_\chi(T_f) = \frac{g_s(T_0)}{g_s(T_f)}\left(\frac{T_0}{T_f}\right)^3 n_\chi(T_f) = \frac{g_{\text{eff}}}{2}\frac{g_s(T_0)}{g_s(T_f)}n_\gamma^0. \tag{10.48}$$

The representative of the hot dark matter is the neutrino. As the number density was discussed in Eq. (8.113) and the possibility of the neutrino being the hot dark matter is small, we will not discuss the hot dark matter any more.

10.3.3
Cold Dark Matter

If the relic particle is already nonrelativistic at the time of freeze-out, that is, if it is a cold dark matter (CDM), we can use Eq. (8.40) and we have [4]

$$Y_\chi(\infty) = \frac{n_\chi(T_f)}{s(T_f)} = \frac{g_{\text{eff}}\left[\dfrac{m_\chi T_f}{2\pi}\right]^{\frac{3}{2}} e^{-\frac{m_\chi}{T_f}}}{g_s(T_f)\dfrac{2\pi^2}{45}T_f^3} = \frac{45}{2\pi^4}\left(\frac{\pi}{8}\right)^{\frac{1}{2}}\frac{g_{\text{eff}}}{g_s(T_f)}x_f^{\frac{3}{2}}e^{-x_f}, \tag{10.49a}$$

$$x_f = m_\chi/T_f \gg 3, \tag{10.49b}$$

$$n^0_{\chi,\text{CDM}} = \frac{s(T_0)}{s(T_f)}n_\chi(T_f) = \frac{s(T_0)}{n^0_\gamma}n^0_\gamma Y_\chi(\infty) = \left(\frac{\pi}{32}\right)^{\frac{1}{2}}\frac{g_{\text{eff}}}{\zeta(3)}\frac{g_s(T_0)}{g_s(T_f)}n^0_\gamma x_f^{\frac{3}{2}}e^{-x_f}. \tag{10.49c}$$

Figure 10.10 shows the numerically obtained solutions for different values of annihilation cross section. Notice that the freeze-out temperature critically depends on the reaction cross section as well as on the mass. For weakly interacting massive particles (WIMPs), x_f is about 20–30 (see discussions after Eq. (10.54)).

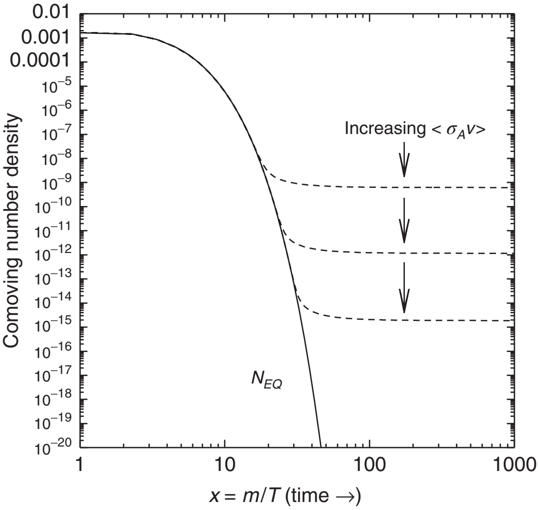


Figure 10.10 Process of the thermal freeze-out of stable, weakly interacting particles in the early universe. As the temperature drops below the particle's mass ($x \gtrsim 1$), the particles are still in thermal equilibrium but their number density is Boltzmann-suppressed. As the universe expands further, eventually these particles are decoupled from other particles and the number density is frozen. The level of frozen density is a function of the annihilation cross section. (Reproduced with permission of [557].)

---

[4] We ignore the chemical potential, that is, consider only the nondegenerate case.

Here we discuss how to observe the nonrelativistic heavy particle, that is, CDM as the relic of the Big Bang. For the CDM to be observable, its cosmic density needs to be of the order 1.

$$\begin{aligned}\Omega_\chi &= \frac{m_\chi n_\chi^0}{\rho_{c0}} = \frac{m_\chi}{\rho_{c0}}\frac{s(T_0)}{s(T_f)}n_\chi(T_f) = \left(\frac{8\pi G_N}{3H_0^2}\right) m_\chi n_\chi(T_f)\frac{g_s(T_0)}{g_s(T_f)}\left(\frac{T_0}{T_f}\right)^3 \\ &\overset{(10.49a)}{=\!=} \frac{8\pi G_N}{3H_0^2}\frac{g_s(T_0)}{g_s(T_f)}\left(\frac{T_0}{T_f}\right)^3 m_\chi \left\{ g_{\text{eff}}\left(\frac{m_\chi T_f}{2\pi}\right)^{\frac{3}{2}} e^{-\frac{m_\chi}{T_f}} \right\} \\ &= \frac{2}{3} g_{\text{eff}} \sqrt{\frac{2}{\pi}} \frac{G_N}{H_0^2}\frac{g_s(T_0)T_0^3}{g_s(T_f)} m_\chi x_f^{\frac{3}{2}} e^{-x_f} \sim 1\end{aligned} \tag{10.50}$$

Equation (10.50) constrains the range of x_f, but this equation alone cannot determine the value of x_f. The freeze-out temperature can be determined by the condition $\Gamma = H$. Using the expression Eq. (8.60) of the Hubble parameter in the radiation-dominated era, it leads to

$$n_\chi(T_f) = \frac{\Gamma}{<\sigma v>} = \frac{H}{<\sigma v>} \overset{(8.60)}{=\!=} \sqrt{\frac{4\pi^3}{45} g_*(T_f)} \frac{T_f^2}{M_{\text{Pl}}} \frac{1}{<\sigma v>} \tag{10.51}$$

where

$$g_*(T) = \sum_B g_B \left(\frac{T_B}{T}\right)^4 + \frac{7}{8}\sum_F g_F \left(\frac{T_F}{T}\right)^4 \tag{10.52}$$

Equating Eq. (10.51) with the number density $\left[\text{content of } \{\cdots\} \text{ in Eq. (10.50)}\right]$, we have an equation to determine the freeze-out temperature.

$$n_\chi(T_f) = g_{\text{eff}} \left[\frac{m_\chi T_f}{2\pi}\right]^{\frac{3}{2}} e^{-\frac{m_\chi}{T_f}} = \sqrt{\frac{4\pi^3}{45} g_*(T_f)} \frac{1}{M_{\text{Pl}}} \frac{T_f^2}{<\sigma v>} \tag{10.53}$$

which leads to

$$x_f = \ln\left[\sqrt{\frac{45}{8}}\frac{g_{\text{eff}}}{2\pi^3} M_{\text{Pl}}\right] + \ln\left(m_\chi <\sigma v>\right) + \frac{1}{2}\ln\frac{x_f}{g_*(T_f)} \tag{10.54}$$

x_f can be iteratively solved, or at least an approximate value can be obtained by neglecting mass dependence of the cross section and with the third term on the r.h.s. replaced with the first approximation of x_f. Further, the requirement that $\Omega_\chi \sim O(1)$ (Eq. (10.50)) constrains the value of x_f approximately $20 \lesssim x_f \lesssim 30$ for $m_\chi \sim 10$–1000 GeV. Substituting the last equality of Eq. (10.53) into Eq. (10.50), the present CDM density can also be rewritten as

$$\Omega_\chi = \frac{m_\chi}{\rho_{c0}}\frac{s(T_0)}{s(T_f)} n_\chi(T_f) \overset{(10.51)}{=\!=} \frac{16\pi^{\frac{5}{2}}}{9\sqrt{5}} \frac{g_s(T_0)\sqrt{g_*(T_f)}T_0^3}{g_s(T_f) M_{\text{Pl}}^3 H_0^2} \frac{m_\chi}{T_f} \frac{1}{<\sigma v>} \tag{10.55}$$

Inserting numbers in Eq. (10.55) and setting $g_*(T_f) = g_s(T_f) = N(T_f)$ (Eq. (8.48)), one gets [834]

$$\Omega_\chi h^2 \simeq 3.4 \times 10^{-38}\,\text{cm}^2 \frac{x_f}{\sqrt{N(T)}} \frac{1}{<\sigma v>} \simeq \frac{0.1\text{pb}}{<\sigma v>} \tag{10.56}$$

where $x_f \simeq 30$ and $N(T) \simeq 100$ is used in obtaining the last equality. This is approximately a mass-independent formula because the value of $x_f/\sqrt{N(T_f)}$ is limited to a narrow range. From Eq. (10.56), one sees that the cross section of the CDM is of the order of weak interaction. Therefore, the CDM should be a WIMP if its mass is substantial, that is, $m_\chi >$ a few gigaelectronvolts.

To get some insight into the mass dependence of the density, let us use an approximate formula valid for the weak interaction in the SM. For simplicity, we also assume it has a spin 1/2. Then the annihilation cross section can be approximated by

$$<\sigma v> \simeq \begin{cases} G_F^2 m_\chi^2 & T \ll m_W \\ G_F^2 \dfrac{m_W^4}{m_\chi^2} & T \gg m_W \end{cases} \tag{10.57}$$

After insertion of the numbers, Eq. (10.55) can be expressed as

$$\Omega_\chi h^2 \simeq 10^{-4} \times \frac{x_f}{\sqrt{N(T_f)}} \left(\frac{G_F^2 m_W^2}{<\sigma v>}\right) \times \begin{cases} \left(\frac{m_W}{m_\chi}\right)^2 & T_f \ll m_W \\ \left(\frac{m_\chi}{m_W}\right)^2 & T_f \gg m_W \end{cases} \tag{10.58}$$

For $T_f \ll m_W$, Eq. (10.54) gives $x_f \sim 20$. Adopting an approximate value of $N(T_f) = N(m_\chi/x_f) \simeq 80$, the requirement of $\Omega_\chi h^2$ to be less than the observed value ($\Omega_{\text{CDM}} h^2 = 0.11$) leads to a lower bound for $m_\chi \gtrsim\sim 6$ GeV. For $T_f >> m_W$, similar arguments lead to an approximate upper bound $m_\chi < 1$ TeV [835]. In other words, the mass of the CDM is limited both from below and from above.[5] The detailed value depends on the model.

10.3.4
Candidates for the Dark Matter

There are several candidates for the dark matter. Cosmological arguments constrain its major component to be nonbaryonic ($\Omega_b \ll \Omega_m$), as can be seen in Table 9.1. However, the baryonic matter may still contribute as subdominant components. Nonbaryonic candidates include heavy neutrinos and neutralinos (generically referred to as *WIMPS*) and axions. Detection methods differ depending on which of the candidates are to look for.

---

[5] The argument was first applied to derive the forbidden region of the neutrino mass (1 MeV $< m_\nu <$ 2 GeV), the so-called Lee–Weinberg bound [836]. There, the neutrino was treated as a massive CDM. The bound was extended later experimentally from absence of $Z \to \nu\bar{\nu}$ with mass $m_\nu < m_Z/2$.

MACHO (massive compact halo object): It is a name given to baryonic dark matter, that is, nonluminous matter in the form of brown dwarfs, neutron stars, and black holes, that float in the galactic halo. When they cross lines connecting nearby stars to the Earth, brightness of the stars increases because of the gravitational lensing effect. In the MACHO searches, time variation of 850 million stars in the Magellanic Clouds were measured. Proof of the lens effect is given by the achromaticity of the brightness change; that is, the effect of the gravitational lensing should be identical over the whole spectral range. Such effects were indeed observed [837], but with too small a frequency to be a major part of the dark matter. Studies revealed [838] that their contribution to the mass of the galactic halo more than 8% is excluded, strengthening the need for nonbaryonic dark matter.

Neutrinos: If there is a neutrino heavier than $\gtrsim 30$ eV but lighter than 1 MeV, it could be a hot dark matter candidate. However, both accelerator experiments $Z^0 \rightarrow \nu\bar{\nu}$ and cosmological arguments excluded the existence of light ($m_\nu \lesssim 1$ MeV) neutrino. However, if heavy neutrinos $m_\nu \gtrsim m_Z/2$ GeV exist, they can be candidates for the CDM. It is a kind of WIMPs and its detection method is very similar to that of the neutralino.

Axions: The axion is very light ($m_a \ll O(1)$ eV). However, the so-called misalignment axion (see Section 7.4.3) is degenerate in the Bose–Einstein condensate state and is classified as one of the CDMs. Negative results of terrestrial experiments and astrophysical observations have constrained the allowed mass range severely. Only a small window ($10^{-6} < m_a < 10^{-3}$) is left. As the axion can couple to two photons, its passage through a strong electromagnetic field yields a photon (referred to as the *Primakov effect*) having the same energy as the axion mass. Axion solar telescopes and cryogenic microwave techniques are utilized for their detection. The method was already described in Section 7.5. The limits by the solar axion searches were given in Figure 7.11 and those by the microcavity were given in Figure 7.13.

WIMPS: WIMPS is an acronym for weakly interacting massive particles. Among the theories that provide the WIMPS, the supersymmetric theory is the favorite of the particle physics community because it has a virtue to provide a means to unify all the forces including the gravity and to solve the hierarchy problem in grand unified theories (GUTs) as well. The coupling strength of the supersymmetry (SUSY) particle with ordinary particles is the same as its SM partner, but as its reaction typically goes through intermediate states involving heavy SUSY particles, its effective strength becomes of the order of the weak interaction. The lightest supersymmetric particle (LSP) is a Majorana particle, a mixture of SUSY partners of the B, W^0 (i.e., the photon and Z), and neutral Higgs bosons:

$$\tilde{\chi}_i\,(i = 1 \sim 4) = a_1\tilde{B} + a_2\tilde{W}^0 + a_3\tilde{H}_d^0 + a_4\tilde{H}_u^0. \tag{10.59}$$

If R-parity of the SUSY is conserved, the lightest particle is stable. The lightest neutralino $\tilde{\chi}_1$ is the prime candidate for the dark matter. Depending on which of the four components are dominant, its reaction rate is different. If $P = \sqrt{a_1^2 + a_2^2} > 0.9$, the neutralino is almost gaugino-like, and if $P < 0.1$ it is almost Higgsino-like. Scans of the parameter space used to show that the neutralino in the mass range 30 GeV∼a few TeV could give a relic density close to the critical density, but the large hardon collider (LHC) experiments placed severe constraints on the mass of LSPs (see Section 10.4.2).

10.4
How to Detect?

10.4.1
Indirect methods

Astrophysical Observation: The neutralinos that are floating in the galaxy halo or those trapped in the core of the Sun can pair-annihilate. The produced pairs, in turn, pair annihilate or decay to various known particles which can be observed as anomalies in otherwise standard spectra.

$$\chi\chi \to t\bar{t},\ b\bar{b},\ WW,\ ZZ,\ \gamma\gamma \xrightarrow{\text{decay}} e^-,\ p,\ d^-,\ \gamma,\ \nu,\cdots \tag{10.60}$$

Many astrophysical detectors, placed on the Earth's surface or mounted on satellites, are trying to detect such anomalies [557, 834, 839, 840].

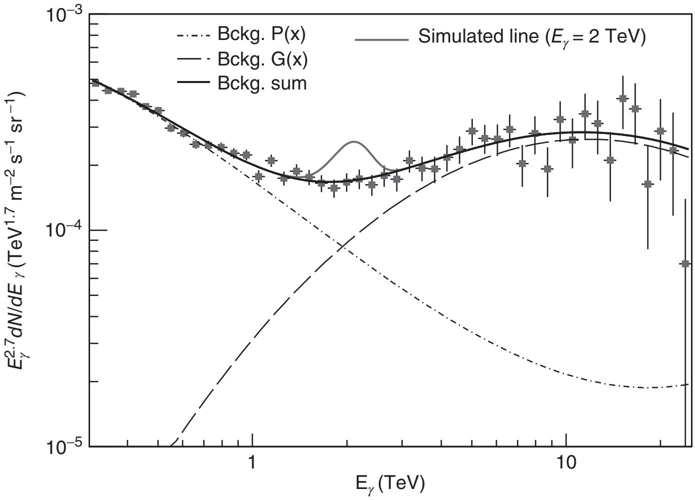


Figure 10.11 γ ray spectrum measured by H.E.S.S. (the High Energy Stereoscopic System) Reconstructed γ-ray flux spectrum of the CGH (central galactic halo) region, using 25 equidistant bins per unit of $\log_{10}(E_\gamma)$. Flux points have been multiplied by $E^{2.7}$. The data consist mostly of hadronic cosmic ray background events, reconstructed using a γ-ray hypothesis. The spectrum is well described by a polynomial parameterization $P(x)$ and a Gaussian function $G(x)$. The two contributions $P(x)$ and $G(x)$ are shown by the dashed-dotted and the dashed curve, respectively. Note that the shape of the Gaussian function $G(x)$ is much broader than the expected monochromatic line feature from DM annihilations. As an example, a hypothetical signal curve (denoted "simulated line" in the legend) shows the expected signal of a line at $E_\gamma = 2$ TeV, which would be detected with a statistical significance of 5σ above the background. (Reproduced with permission of [841].)

An example of the γ ray data coming from the central part of the Milky Way halo region collected by the H.E.S.S. instrument [841] is shown in Figure 10.11. The covered energy region is 500 GeV–25 TeV. The spectra consist mostly of γ-ray-like cosmic ray background events (and a fraction of 10% of electrons). The flux points are multiplied by $E^{2.7}$. These spectra are well described by the empirical parameterization using polynomials (dashed-dotted) in powers of E_γ and a Gaussian (dashed). The shape of the Gaussian function is much broader than the expected monochromatic line feature from DM annihilations. An example of expected 5σ DM signal of a line at $E_\gamma = 2$ TeV is shown as a small peak (denoted "simulated line"). For a DM particle mass of 1 TeV, limits on the velocity-averaged DM annihilation cross section $<\sigma v>_{\chi\chi\to\gamma\gamma}$ reach $\sim 10^{-27}\mathrm{cm}^3\mathrm{s}^{-1}$, based on the Einasto parameterization [842] of the Galactic DM halo density profile.

Underground Observation: Here we will briefly describe searches for high-energy neutrinos created in the Sun by pair annihilation of the captured neutralinos [834, 843]. The WIMPs can be gravitationally trapped inside the Sun and the Earth. It happens when the WIMP is scattered and its final states have velocities below the escape velocity. The annihilation rates in the Sun and the Earth are estimated using the Standard Halo Model (SHM), which is modeled as a smooth, spherically symmetric density component with a nonrotating Gaussian velocity distribution. The trapped WIMPS annihilate each other and emit high-energy neutrinos [844–847]. On the Earth, these neutrinos produce muons via charged current interactions with nuclei: $\nu_\mu + A \to \mu + X$. Large detectors (Super-Kamiokande [848], AMANDA [849], IceCube [850], etc.) placed deep underground or in the Antarctic area are used to detect those muons. So far, no sign of neutralino annihilation in the Sun have been observed. Figure 10.12 shows the result of the IceCube together with other large detectors [850].

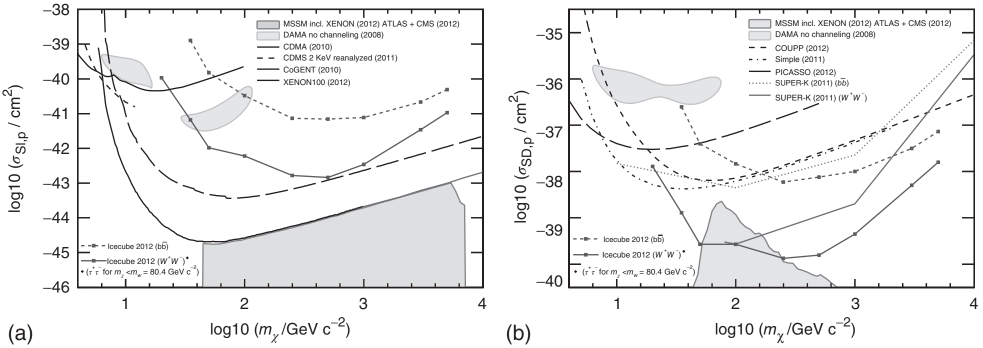


Figure 10.12 90% CL upper limits on (a) spin-independent and (b) spin-dependent WIMP-nucleon cross sections for hard and soft annihilation channels over a range of WIMP masses. The shaded region represents the allowed MSSM parameter space (MSSM-25) [851]. Results from other experiments are shown for comparison. (Reproduced with permission of [850].)

10.4.2
Production by Accelerators

The WIMPS, if they have right mass and coupling, can be produced by accelerators. A typical WIMP is the LSP or LKP(lightest KK particle) but there are other theories that predict similar particles. As the SUSY LSP is stable and interacts only weakly with the SM particles, it will appear as a missing energy tagged by a jet or photon from the initial- or final-state radiation [852, 853]. Previous searches at the large electron–positron collider (LEP) or Tevatron did not find any, and the lower mass limit of ∼47 GeV was obtained [7]. At the LHC, the mass range is greatly extended.

Model-independent searches were typically based on the effective theory approach with contact interactions of the form given in Table 10.2. This is because the SUSY particles are typically pair-produced by the R-parity requirement. The effective coupling constant is parameterized by the mass scale M_*. Within this framework, interactions of SM and WIMPS are described by only two parameters, namely M_* and the WIMP mass m_χ.

The signal would be an excess of events having the topology "mono jet $+\not{E}_T$ (missing transverse energy)" over those of SM. The SM backgrounds come form Z/W+jets, $t\bar{t}$+ single top, and dibosons. No signals were found at LHC. An example of the event production rate as a function of $\not{E}_T$ was already given in Figure 6.18. The data was taken by the CMS group [584] in the search for the graviton in LED (large extra dimension) theories and the WIMP as well.

Figure 10.13 shows another set of data taken by the ATLAS group. Upper limits on the (a) spin-independent and (b) spin-dependent WIMP-nucleon cross sections were obtained [585].

Table 10.2 Examples of effective interactions coupling Dirac fermion WIMPS to SM quarks or gluons following the formalism of [854]. The tensor operator describes a magnetic coupling. The factor of the coupling constant α_s in the definition of the operator accounts for the interaction induced at the one-loop level. $G_{\mu\nu}$ is the color field strength.

Name	Reaction	Type	Operator
D1	qq→ $\bar{\chi}\chi$	Scalar	$\frac{m_q}{M_*^3}(\bar{\chi}\chi)\,(\bar{q}\,q)$
D5	qq→ $\bar{\chi}\chi$	Vector	$\frac{1}{M_*^2}(\bar{\chi}\gamma^\mu\chi)\,(\bar{q}\gamma_\mu q)$
D8	qq→ $\bar{\chi}\chi$	Axial-vector	$\frac{1}{M_*^2}(\bar{\chi}\gamma^\mu\gamma^5\chi)\,(\bar{q}\gamma_\mu\gamma^5 q)$
D9	qq→ $\bar{\chi}\chi$	Tensor	$\frac{1}{M_*^2}(\bar{\chi}\sigma^{\mu\nu}\chi)\,(\bar{q}\sigma_{\mu\nu}q)$
D11	gg → $\bar{\chi}\chi$	Scalar	$\frac{\alpha_s^2}{4M_*^3}(\bar{\chi}\chi)(G^a_{\mu\nu})^2$

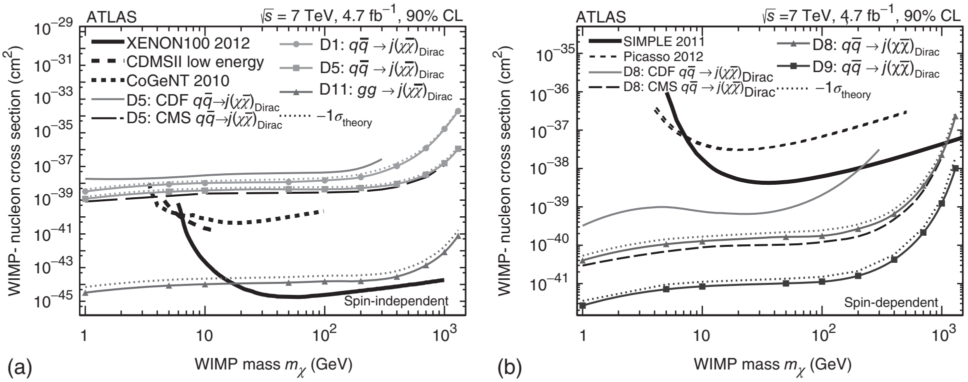


Figure 10.13 (a) Inferred 90% CL ATLAS limits on spin-independent (a) and spin-dependent (b) WIMP-nucleon scattering. Cross sections are shown as a function of the WIMP mass m_χ. In all cases, the thick solid lines are the observed limits excluding theoretical uncertainties; the observed limits corresponding to the WIMP-parton cros section are shown as thin dotted lines. The latter limits are conservative because they also include theoretical uncertainties. The ATLAS limits for operators involving quarks are for the four light flavors assuming equal coupling strengths for all quark flavors to the WIMPs (See Table 10.2 and [854] for the meaning of D_is.). For comparison, the 90% CL limits for spin-independent cross sections from the XENON100 [855], CDMSII [856], CoGeNT [857], CDF [858], and CMS [584, 859] experiments and spin-dependent cross sections from the SIMPLE [860], Picasso [861], CDF [858], and CMS [584, 859] experiments are shown. (Reproduced with permission of [585].)

10.4.3
WIMPS Wind

The dark matter resides in the halo of the galaxies [834, 840, 862]. Their average velocity is zero. As the Earth is circulating in the galactic halo, the WIMPs have a relative velocity to the Earth. That is, the WIMPs are blowing as a wind on the Earth. They do not interact electromagnetically, or strongly, and hence receive no forces from ordinary matter except gravity. Therefore, the existence of the dark matter in the galactic halo is of cosmological origin. Catching the halo dark matter is probably the most promising non-accelerator method which we will describe in detail.

Mass Density The distribution of the dark matter in the galactic halo is assumed to be spherical obeying isothermal Maxwell distribution $f(\boldsymbol{v}) \sim e^{-v^2/v_0^2}$. That is, the average velocity is zero but its dispersion is not. Two models exist for the number distribution to explain the observed rotational velocity of the Milky Way galaxy (see Figure 10.14). One is flat and the other has a cusp ($r \propto r^\alpha$, where r is the distance from the galactic center).

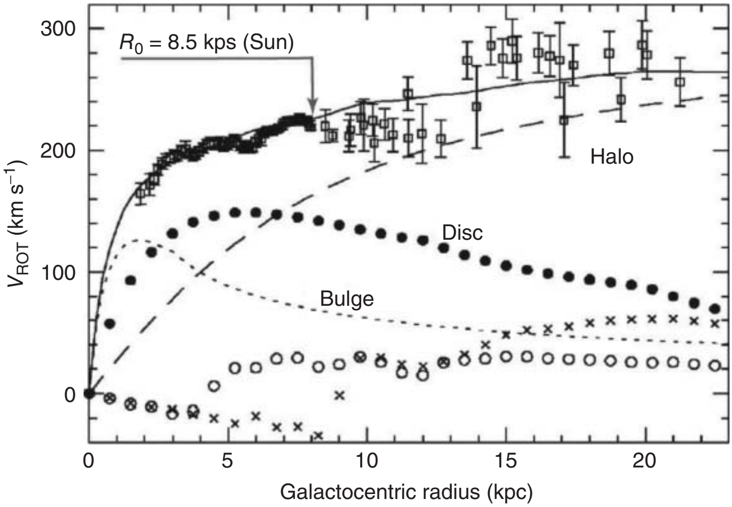


Figure 10.14 Velocity curve of the Milky Way. Contributions of various components are denoted. (Reproduced with permission of [863].)

$$\rho_{\mathrm{DM}} = \frac{\rho_0}{1+\frac{r^2}{r_0^2}} \quad \text{or} \quad \frac{\rho_s}{\left(\frac{r}{r_s}\right)^{\alpha}\left(1+\frac{r}{r_s}\right)^{3-\alpha}}, \quad \alpha = 0 \sim 1$$

$$f(\mathbf{v}) = \frac{1}{(\pi v_0^2)^{\frac{3}{2}}} e^{-\frac{v^2}{v_0^2}} \quad \rightarrow \quad f(|\mathbf{v}|) = \frac{4\pi v^2}{(\pi v_0^2)^{\frac{3}{2}}} e^{-\frac{v^2}{v_0^2}} \tag{10.61}$$

$$\rho_0 \simeq 0.3\,\mathrm{GeV\,cm^{-3}}, \quad v_0 \simeq 220\,\mathrm{km\,s^{-1}}, \quad \rightarrow \quad v_{\mathrm{rms}} = \sqrt{\frac{3}{2}} v_0 = 270\,\mathrm{km\,s^{-1}}$$ [6]

The energy density of the DM is estimated to be 0.3 GeV cm^{-3} [676, 815, 864]. Therefore, if one assumes $m_{\mathrm{WIMP}} \sim 100$ GeV, then its number density is calculated to be $\sim 3000/m^3$, and about $6 \times 10^4/\mathrm{cm}^2$ WIMPS are passing through our body every second.

Detection Rate Denoting the DM's number density n_0 at the Earth's position ($r = R_0 = 8.5$ kpc), its velocity distribution function relative to the detector $f(\mathbf{v})$, and the scattering cross section $\sigma_{\chi N}$, the scattering rate R of DM per unit volume is given by [840]

$$\begin{aligned} R &= n_0 \int d\mathbf{v} f(\mathbf{v})\, v \sigma_{\chi N} n_{\mathrm{tgt}}, \\ n_0 &= \frac{\rho_{\mathrm{DM}}}{m_\chi}, \qquad \int d\mathbf{v} f(\mathbf{v}) = 1 \end{aligned} \tag{10.62}$$

---

[6] Compare v_0 with the escape velocity of the Milky Way which is ~550 km. Note the circulating velocity of the solar system in the galaxy is ~244 km s^{-1}.

where n_{tgt} is the number density of the target material in the detector, and ρ_{DM} and m_χ are the energy density and mass of the DM, respectively. If the DM is the neutralino in SUSY, the scattering rate is calculable as a function of its mass. The theoretically allowed range of the mass used to be 100–1000 GeV. Recent LHC experiments raised the lower limit up to ~800 GeV (see discussions in Section 5.7).

Cross Section The DM is electrically neutral but it can be detected by observing the scattered particle in the detector. Detection of the DM is hard because it not only interacts weakly with ordinary matter but also gives only a small recoil energy to the detector particle. Assuming the DM's mass to be 100 GeV, its kinetic energy is given by

$$\text{KE} = \frac{1}{2} m_\chi v^2 \sim \frac{1}{2} \times 100\,\text{GeV} \times \left(220\,\text{km s}^{-1}\right)^2 \simeq 27\,\text{keV}. \tag{10.63}$$

Consequently, we are talking about very low energy elastic scattering, with isotropic angular distribution. The cross section is also proportional to the reduced mass of the DM and the target particle.

Problem 10.2 *Consider the low-energy ($E \ll m$) scattering cross section due to the Yukawa-type potential ($V = \frac{g^2}{4\pi} e^{-mr}$). Show that, in the Born approximation, the cross section is isotropic in the CM system and that the cross section is proportional to the square of the reduced mass [$\mu = m_1 m_2/(m_1 + m_2)$].*

We denote the cross section of the DM with a nucleon and a nucleus (having mass $m_A \simeq 0.932A$ with A the atomic number) as $\sigma_{\chi N}$ and $\sigma_{\chi A}$, respectively. Two kinds of scatterings exist, one denoted as SI (spin independent), which does not depend on the spin of the participating particles, and the other as SD, which is spin dependent. If one assumes the DM to be the SUSY particle, SI and SD processes include the diagrams depicted in Figure 10.15.

Assuming $m_\chi = 100$ GeV, $v \simeq 10^{-3}$, its de Broglie wavelength becomes $\lambda \sim 3$ fm (1fm = 10^{-15} m). This is comparable with the nuclear size. If the de Broglie wavelength is larger than the nuclear size, target nucleons react coherently and the

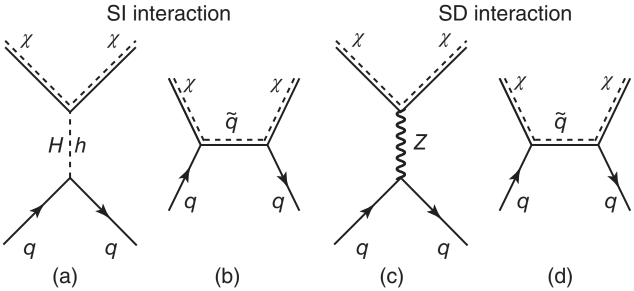


Figure 10.15 Feynman diagrams for the neutralino (χ)–quark scattering. Double lines of dotted and real indicate the neutralino (χ), which is a Majorana particle with spin 1/2.

cross section scales as A^2. Consequently

$$\begin{aligned}
&\sigma_{\chi A} = \sigma_A^{\mathrm{SI}} + \sigma_A^{\mathrm{SD}} \\
&\sigma_A^{\mathrm{SI}} \sim A^2 \frac{\mu_A^2}{\mu_N^2}\sigma_N^{\mathrm{SI}}, \qquad \sigma_A^{\mathrm{SD}} \simeq \frac{3}{4}\lambda^2 J(J+1)\frac{\mu_A^2}{\mu_N^2}\sigma_N^{\mathrm{SD}} \\
&\mu_A = \frac{m_\chi m_A}{m_\chi + m_A}
\end{aligned} \tag{10.64}$$

where λ is the Lande factor, A is the atomic number, and J is the total angular momentum of the target nuclei. The cross section becomes maximum around $m_\chi \sim m_A$. If $m_\chi \sim m_A \gg m_N$, $\mu_A \simeq Am_N$, which means $\sigma^{\mathrm{SI}} \propto A^4$, and the counting rate per unit mass is proportional to $\sim A^3$. These considerations suggest the advantage of using a material with large A.

The ratio of the spin-dependent cross section and the scalar cross section ($\sigma^{\mathrm{SD}}/\sigma^{\mathrm{SI}}$) depends on the model. It is typically as large as $10^2 \sim 10^4$, but experimentally the counting rate of the SI target being proportional to $\sim A^3$, SI has better sensitivity in general.

Recoil Energy When the energy of the incident DM ($E_i = m_\chi v_i^2/2$) is given, the recoil energy E_R of the target particle (mass m_A) in the laboratory system is related to the scattered angle ϕ in LAB and θ^* in the center of mass system by

$$E_R = \frac{E_i}{2} r(1 - \cos\theta^*), \qquad r = \frac{4m_\chi m_A}{(m_\chi + m_A)^2}, \quad \frac{1}{2}(1 - \cos\theta^*) = \cos^2\phi. \tag{10.65}$$

Consequently, $dE_R/d\Omega_{\mathrm{CM}}$ is a constant. Generally, two-body elastic scattering at the low energy is isotropic and it is also assumed here. Then the recoil energy spectrum is uniform in the range $0 \le E_R \le E_i r$. That is, when v_i is fixed, the total cross section σ_0 is given by

$$\frac{d\sigma}{dE_R} = \frac{d\sigma}{d\Omega_{\mathrm{CM}}}\left(\frac{dE_R}{d\Omega_{\mathrm{CM}}}\right)^{-1} = \text{const.} = \frac{\sigma_0}{E_i r}, \qquad 0 \le E_R \le E_i r \tag{10.66}$$

Problem 10.3 *Derive Eq. (10.65).*

The event rate N as a function of E_R is given as the average over the velocity distribution.

$$\frac{dN}{dE_R} = \left\langle n(\mathbf{v}) v \frac{d\sigma}{dE_R} \right\rangle n_{\mathrm{tgt}} \tag{10.67}$$

where $n(\mathbf{v}) = n_0 f(\mathbf{v})$ with the average defined by $<\cdots>= \int d^3v f(\mathbf{v})(\cdots)$. Substituting Eq. (10.61) (10.66) into Eq. (10.67), we obtain

$$\begin{aligned}
&\frac{dN}{dE_R} = \int_{v_{\min}}^{v_{\max}} d^3v\, n_0 n_{\mathrm{tgt}} v f(\mathbf{v}) \frac{\sigma_0}{E_i r} = \frac{n_0 \sigma_0 v_0^2 n_{\mathrm{tgt}}}{E_0 r} F^2(E_R) \int_{v_{\min}}^{v_{\max}} d^3v \frac{f(\mathbf{v})}{v} \\
&n_0 = \frac{\rho_{\mathrm{DM}}}{m_\chi}, \qquad E_0 = \frac{m_\chi}{2} v_0^2
\end{aligned} \tag{10.68}$$

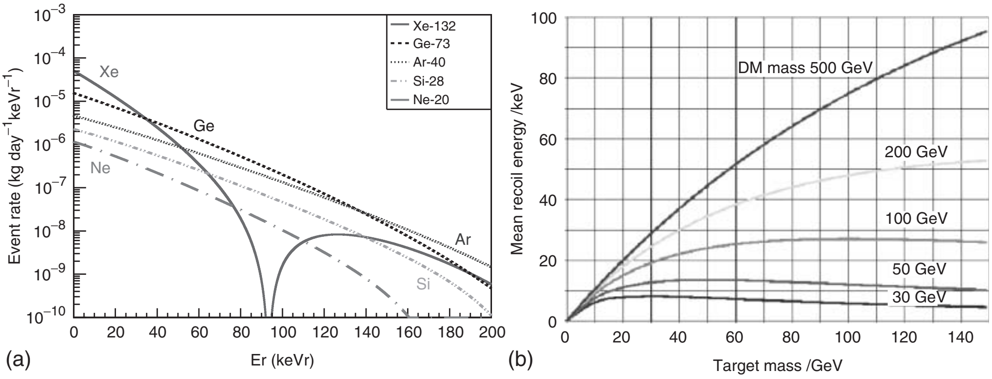


Figure 10.16 (a) Differential nuclear recoil spectra for Xe and Ar and other nuclei, calculated from Eq. (10.70) for a WIMP mass of 100 GeV. The dip in the Xe curve is the nuclear form factor effect. (Reproduced with permission of [866, 867].). (b) Average recoil energy ($< E_R > = E_0 r$) as a function of target mass m_A. Differential WIMP recoil spectrum for WIMP mass of 100 GeV and WIMP-nucleon cross section $\sigma = 10^{43}\, cm^2$ [840].

where the nuclear form factor $F(E_R) < 1$ was added to correct deviation from coherent scattering (see the next section). $v_{\text{min}} = \sqrt{2E_R/m_\chi r}$ is the minimum velocity to satisfy Eq. (10.65) for given E_R, and $v_{\text{max}} \simeq 550$ km s^{-1} is the escape velocity of the DM from the galaxy. Extending $v_{\text{max}} \to \infty$ gives little modification, and substituting

$$\int_{v_{\text{min}}}^{\infty} d^3v \frac{f(\mathbf{v})}{v} = \frac{2}{\sqrt{\pi} v_0} e^{-\frac{E_R}{E_0 r}} \tag{10.69}$$

into Eq. (10.68) we obtain [840, 865]

$$\begin{aligned}
\frac{dN}{dE_R} &= \frac{R_0}{E_0 r} F^2(E_R) e^{-\frac{E_R}{E_0 r}}, \\
R_0 &= \int_0^{\infty} dE_R \frac{dN}{dE_R} = \frac{2}{\sqrt{\pi}} n_0 v_0 \sigma_0 n_{\text{tgt}}, \\
< E_R > &= \frac{1}{R_0} \int dE_R\, E_R \frac{dN}{dE_R} = E_0 r \sim 27\,\text{keV} \quad \text{for } m_\chi = m_A = 100\,\text{GeV}
\end{aligned} \tag{10.70}$$

The spectrum and mean value of the recoil energy are given in Figure 10.16.

As the recoil energy spectrum is very similar to that of detector noise, construction of a detector with a small energy threshold and small noise is required.

The above formula was calculated assuming that the Earth is at rest and setting the escape velocity as infinite. The following approximate formula takes into account of the Earth's motion [840, 865], which is described later [Eq. (10.78)]:

$$\frac{dN}{dE_R} = c_1 \frac{R_0}{E_0 r} F^2(E_R) e^{-\frac{c_2 E_R}{E_0 r}} \quad [\text{counts}^{-1}\ \text{keV}^{-1}\ \text{kg}^{-1}\ \text{day}^{-1}], \tag{10.71a}$$

$$R_0 = \frac{433}{m_\chi(\text{GeV}) m_A(\text{GeV})} \left(\frac{\sigma_0}{1\text{pb}}\right) \left(\frac{\rho_{\text{DM}}}{0.3\,\text{GeV cm}^{-3}}\right) \times \left(\frac{v_0}{230\,\text{km s}^{-1}}\right) [\text{counts}^{-1}\ \text{kg}^{-1}\ \text{day}^{-1}], \tag{10.71b}$$

$$c_1 = 0.751, \quad c_2 = 0.561, \quad E_0 : \text{ kinetic energy of DM.} \tag{10.71c}$$

In deriving R_0, we used $n_{\text{tgt}} \propto N_A/A$, $A = 0.932 m_A(in\ \text{GeV})$, N_{A} = Avogadro number. In the above expression, c_1, c_2 are constants that change monthly due to revolution around the Sun, but only the average over a year is given because the monthly change is small (~3 %). We will describe the annual modulation later.

Nuclear Form Factor De Broglie wavelength of the recoil momentum is of the order of

$$ƛ_R = \frac{1}{q} = \frac{1}{\sqrt{2 m_A E_R}} \simeq \frac{140}{\sqrt{A E_R(\text{keV})}}\,\text{fm} \qquad : 1\text{fm} = 10^{-15}\text{m}. \tag{10.72}$$

On the other hand, the nucleus spreads over a distance

$$r_A \simeq a A^{\frac{1}{3}} + b, \qquad a \sim O(1\text{fm}),\ b \lesssim O(1\text{fm}). \tag{10.73}$$

Therefore, when the recoil energy becomes

$$q r_A = \frac{r_A}{ƛ_R} \simeq \frac{A^{\frac{5}{6}} \sqrt{E_R(\text{keV})}}{140} \gtrsim 1 \tag{10.74}$$

each nucleon inside the nucleus A do not interfere coherently, and the cross section becomes proportional to A rather than A^2, which modifies the formula Eq. (10.64). The effect can be taken into account by multiplying the nuclear form factor

$$\sigma_A \to \sigma_A F(q r_A)^2 \tag{10.75a}$$

$$F(E_R) \equiv F(q r_A) = \int d^3x\, e^{i\mathbf{q}\cdot\mathbf{r}} \rho(\mathbf{r}). \tag{10.75b}$$

$\rho(\mathbf{r})$ is the nuclear density function and can be measured by electromagnetic scattering (see Section 17.3 of [1]). Typically, we may use [840]

$$F(q r_A) = 3\frac{j_1(q r_A)}{q r_A} e^{-(qs)^2} = 3\frac{\sin(q r_A) - q r_A \cos(q r_A)}{(q r_A)^3} e^{-(qs)^2} \tag{10.76}$$

which is the Fourier transform of

$$\rho(\mathbf{r}) = \int \rho_0(\mathbf{r}')\rho_1(\mathbf{r} - \mathbf{r}') d^3\mathbf{r}' \tag{10.77a}$$

$$\rho_0(r) = \text{const.} \quad \text{for } r^2 \le r_A^2 = (1.2 A^{\frac{1}{3}})^2 - 5s^2, \quad s \simeq 1\text{ fm} \tag{10.77b}$$

$$\rho_1(\mathbf{r}) = \exp\left[-\frac{1}{2}\left(\frac{r}{s}\right)^2\right]. \tag{10.77c}$$

The distribution represents a blurred sphere. It is flat to $r \leq r_A$ and then attenuates exponentially with surface thickness $s \sim 1$ fm. Figure 10.16a includes the nuclear form factor effect, as is evident from the dip in the Xe curve. In the following discussion, however, we assume $F^2(E_R) = 1$

Annual Modulation So far we assumed that the Earth is at rest. In reality, it revolves around the Sun with a velocity 30 km s^{-1} and the Sun rotates around the galactic center with a velocity 244 km s^{-1}. The net velocity v_E is almost the same as the dispersion of the DM. The Earth's velocity in the direction of the Sun's motion can be approximated by [840]

$$\mathbf{v}_E = 235\ \text{km} + 30 \cos\gamma \cos(2\pi y)\ \text{km s}^{-1}, \qquad \gamma \simeq \frac{\pi}{3} \tag{10.78}$$

with y the fractional year starting June 2 (Figure 10.17a). The counting rate can be obtained by replacing $f(\mathbf{v}) \to f(\mathbf{v} + \mathbf{v}_E)$. An analytical formula is possible, but practically we can use

$$\frac{dN}{dE_R} = \frac{R_0}{E_0 r} F^2(E_R) e^{-\frac{E_R}{E_0 r}} \quad \to \quad c_1 \frac{R_0}{E_0 r} F^2(E_R) e^{-\frac{c_2 E_R}{E_0 r}}\,. \tag{10.79}$$

As stated before, c_1, c_2 change monthly, but most of the time they are approximated by Eq. (10.71c). The annual modulation is small, with the daily rate changing at most by 3–5% (Figure 10.17b).

If one can construct a direction-sensitive detector, one can measure a strong daily forward/backward asymmetry of the nuclear recoil direction (Figure 10.18). Current R&D efforts are on a time projection chamber (TPC) type gaseous detector and anisotropic-response scintillators (stillbene etc.).

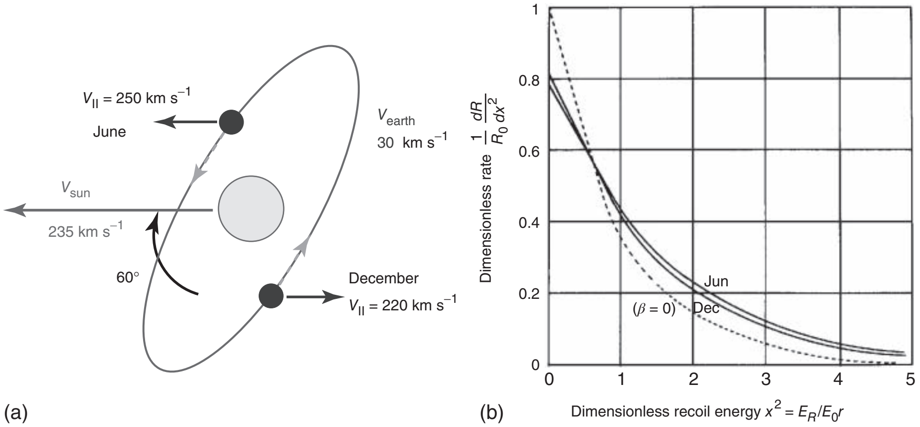


Figure 10.17 (a) Earth's motion in the galaxy. The z-axis is along the Sun's motion, x is radial and away from the galactic center, y is perpendicular to the galactic plane. (b) The annual modulation due to revolution around the Sun is at most a few percent. $\beta = 0$ is the counting rate for the Earth at rest. (Reproduced with permission of [840, 865].)

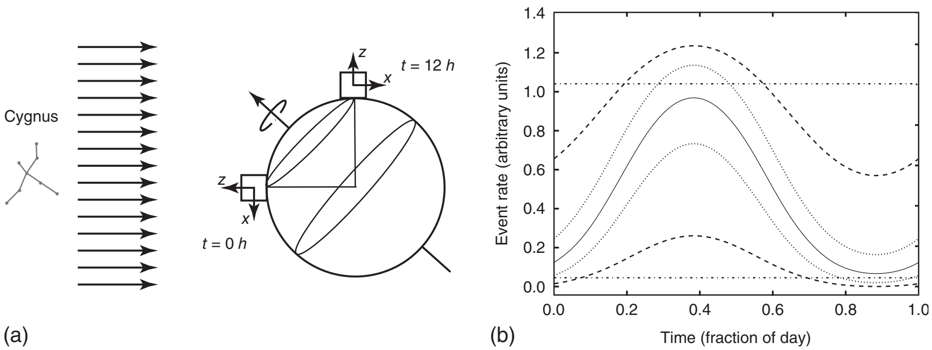


Figure 10.18 (a) Daily rotation of the Earth introduces a modulation in recoil angle, as measured in the laboratory frame. (b) Magnitude of the daily modulation for seven lab-fixed directions, specified as angles with respect to the Earth's equatorial plane. The solid line corresponds to zero degrees, and the dotted, dashed, and dash-dot lines correspond to ±18°, ±54°, and ±90°, with negative angles falling above the zero degree line and positive angles below. The ±90° directions are co-aligned with the Earth's rotation axis and therefore exhibit no daily modulation. This calculation assumes a WIMP mass of 100 GeV and CS_2 target gas. (Reproduced with permission of [868].)

The annual and/or daily modulation would provide a strong ground for claiming evidences for detection of the dark matter. One needs high statistics to detect either of them.

10.5
Searches for DMs in the Halo

10.5.1
General

Signals given by DM are very small (1 ∼ 100 keV). Furthermore, visible energy is even smaller.

$$E_{\text{vis}} = QE_R \tag{10.80}$$

where the quench factor ($Q < 1$) is given as the ratio of the ionization efficiency by the nucleon to that by the electron with the same kinetic energy. A DM detector must have energy threshold as low as possible (∼ a few keV), small noise, and stable operation with good reproducibility. Suppose the DM exists and its predicted rate is given by the dashed line in Figure 10.19.

The actual data (signal + noise) obtained by an unsatisfactory detector may look like the black line. Unless one is sure that the data exceeds the noise level in the neighborhood of $E \simeq E_R$ with supporting evidences, one treats the data with the assumption that there was no sign of positive signals. In this case, the data is treated statistically to give an upper limit of the cross section. For instance,

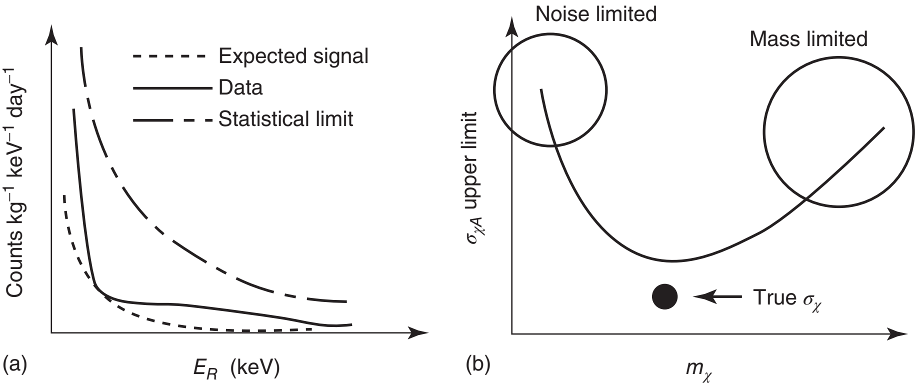


Figure 10.19 Display of (a) analysis method of experimental data and (b) its result. When the noise component dominates, only the upper bound of the cross section (and resultant lower bound on mass of DM) can be given. The true value of the DM mass and cross section is given by the dashed line in (a) and a black circle in (b).

the 95% upper limit line is given by the dot-dashed line in Figure 10.19a. Using theoretical expressions, one can draw a line for the upper limit of the cross section as a function of the DM mass (like the one in Figure 10.19b). The statistical upper limit of the cross section is translated to a lower limit of the dark matter mass. The limit becomes worse at small m_χ due to the detector noise and at large mass due to mass dependence of the counting rate.

From these consideration, one sees that the detector efficiency depends critically on how much one can reduce the threshold energy. The energy spectrum of the recoiled nuclei is exponential, which is similar to that of the background noise. Depending on the cross section, the counting rate of the DM is about $1\text{–}10^{-(5-8)}$ kg day^{-1}, which is usually much weaker than the background noise rate inherent to the detector material or environments. The DM signal depends only on the recoil energy, and no coincidence can be used to reduce the noise. It is essential to reduce the amount of background. Special care is needed for the design of the detector. Some reduction of the noise can be achieved by going deep underground[7] and selecting the detector material having small contamination of radioisotopes. Purification of the material may also be needed. Further reduction of noise can be obtained by using passive and/or active shields to cover the outer surface of the detector. The conventional method is to limit the fiducial volume of event selections. To use materials with different atomic numbers and to observe the A dependence is another way. If enough statistics is available, looking at annual modulation is effective. If one can construct a direction-sensitive detector, daily modulation can be observed. The annual or daily modulation is probably needed for true identification of the DM. A typical detector uses a combination of those mentioned above.

Some exemplary means to detect the signal are to use

---

7) Effects of going underground was shown in Figure 2.20.

(1) ionization loss (electric signal):	Ge,Si,Xe;
(2) scintillation light:	NaI,Xe,$CaWO_4$, BGO;
(3) thermal heat (bolometer):	Ge, Al_2O_3, LiF;
(4) combination of (1)–(3)	

The combination includes (a) ionization and heat (Ge,Si), (b) light and heat ($CaWO_4$, BGO), ionization loss and light (Xe), and so on.

In the following, we discuss two examples of the direct DM detector. One uses a cryogenic bolometer and the other a large mass of Xe.

10.5.2
Bolometer

The liberated energy, whether ionization loss or scintillation light, dissipates eventually as vibration, that is, heat. At the ultracold temperature, the specific heat decreases in proportion to T^3 8) and it is possible to measure small energies around $\lesssim 10$ keV. The energy of a single phonon (quantum of the vibration) is small (∼ meV) but the statistics is so large that the resolution becomes excellent. Figure 10.20a shows a schematic of the bolometer, and Figure 10.20b is a tower of the germanium detectors used by the CDMS group [869].

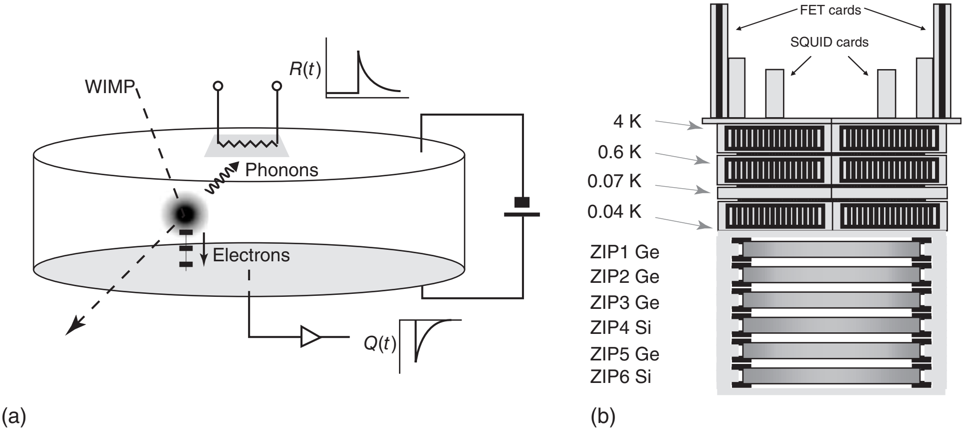


Figure 10.20 (a) Schematic of a Ge bolometer module. Heat is converted to electric signal using a thermistor or SQUID (superconducting quantum interference device). Ionization loss between the electrodes is measured too. (b) Diagram of the CDMS tower. The tower has four temperature stages (4 K, 600 mK, 70 mK, and base). The cold electronics for the ionization channels (FETs) and phonon channels (SQUIDs) are mounted on the top of the tower, which has surfaces at 4 K and 600 mK. The four Ge modules and two Si modules with darker shade are arranged in the bottom portion of the tower with 2-mm separation and no intervening material. (Reproduced with permission of [869].)

---

8) The specific heat at the ultracold temperature is expressed by $C = (12\pi^4 N k_B/5)(T/\Theta_D)^3$, where Θ_D is the Debye temperature.

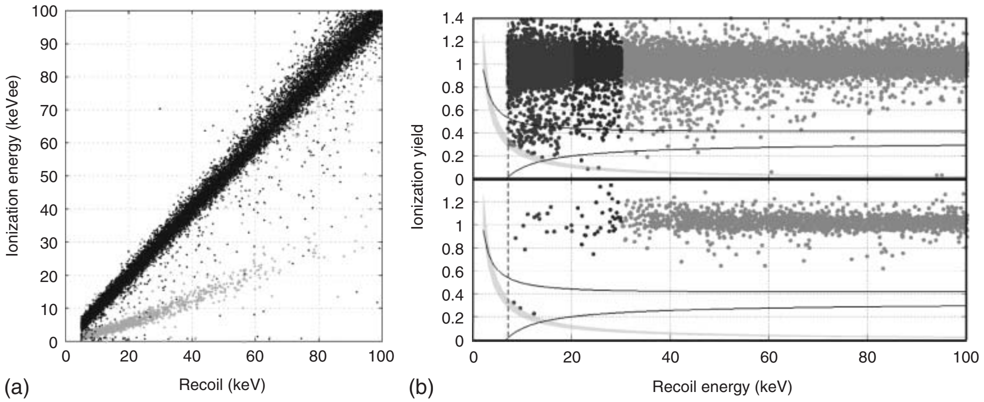


Figure 10.21 (a) Calibration of the phonon signal to the ionization signal for the CDMS detector. (b) Ionization yield versus recoil energy in all detectors for events passing all signal criteria except (top) and including (bottom) the phonon timing criterion. The curved black lines indicate the signal region (-1.8σ and $+1.2\sigma$ from the mean nuclear recoil yield) between 7 and 100 keV recoil energies, while the gray band shows the range of charge thresholds. Electron recoils in the detector bulk have yield near unity. (Reproduced with permission of [869–871].)

The major advantage of the bolometer is its high efficiency of energy detection with a quenching factor close to 1 ($Q \simeq 1$) for the gamma ray and $Q \sim 0.3$ for the neutron. By taking the ratio of the ionization loss to the thermal energy, the two types of the energy loss can be separated with high efficiency.

CDMS Figure 10.21a shows a calibration data of a bolometer used by the CDMS group which shows two distinct responses for the photon and the neutron with a resolution of ~5% at 10 keV.

Figure 10.21b shows the actual data to look for the WIMPS. From the absence of signals associated with nuclear recoils, an upper limit of the DM detection signal was obtained. Bounds for the neutralino mass will be given later in Figure 10.24.

10.5.3
Xe Detector

A different approach for the detection of the cold dark matter using liquid Xenon (LXe) is given here. Although the liquid Xe has a low Q-value (~0.25), it has many advantages.

1) In addition to ionization loss, it also emits a large amount of scintillation light. Moreover, two kinds of scintillation light are emitted, enabling event-by-event particle identification.
2) Its large atomic number ($A = 130$) helps to enhance the coherent cross section. In addition, high density ($\rho = 3\,\mathrm{g\,cm^{-3}}$) and high Z (=54) make

it easy to construct a large size detector with self-shielding ability against electromagnetically interacting particles.

3) Purification is relatively easy, which helps to eliminate long-lived environmental isotopes ^{87}Kr, ^{39}Ar, and so on, and to reduce noises. Two-phase (liquid and gas) operation is possible, reducing the noise further.

Disadvantages are its low *Q* value (~0.25) and need of cooling (−100°). It is also relatively expensive.

The principle of light emissions by the Xe is illustrated in Figure 10.22a and a method of extracting two signals with distinct characteristics at different timing is given in Figure 10.22b,c.

(1) **Xe detector as a scintillation counter**: Xe is ionized by scattering with electrons or nuclei. When the electric field does not exist or is weak, the ionized Xe and the electron recombine immediately, and hence the signals by direct excitation and those through recombination almost overlap (signal S1). In this case, ionization by nuclear recoil is larger than that by electron recoil. Besides, the recombination speed of ionized Xe by electrons ($\sim 40 \times 10^{-9}$ s) is much lower than that by heavy

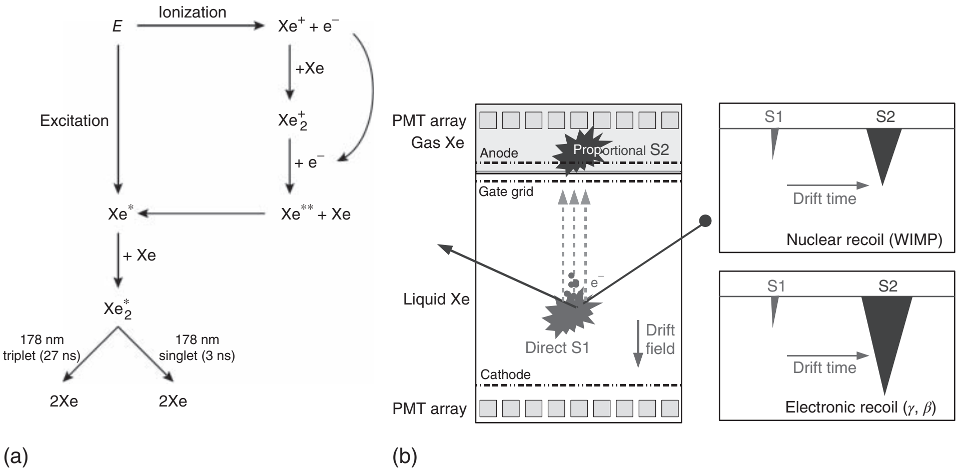

Ionization
E
Xe⁺ + e⁻
+Xe
Excitation
Xe₂⁺
+ e⁻
Xe*
Xe** + Xe
+ Xe
Xe₂*
178 nm triplet (27 ns)
178 nm singlet (3 ns)
2Xe
2Xe
PMT array
Gas Xe
Anode
Proportional S2
Gate grid
Liquid Xe
e⁻
Direct S1
Drift field
Cathode
PMT array
S1
S2
Drift time
Nuclear recoil (WIMP)
S1
S2
Drift time
Electronic recoil (γ, β)
(a)
(b)

Figure 10.22 (a) Schematic of light emission mechanism by Xe. (b) Schematic of the XENON two-phase (liquid–gas) time projection chamber (TPC). A particle interaction in the liquid xenon (LXe) produces direct scintillation photons and ionization electrons. An electric field is applied across the liquid Xe volume with appropriate potentials on a series of electrodes, drifting the ionization electrons away from the interaction site. Electrons that reach the liquid–gas interface are extracted into the Xe gas, where the process of proportional scintillation (electroluminescence) takes place. Both the direct (S1) and the proportional (S2) scintillation light, with 178 nm wavelength, are detected by photomultiplier tubes (PMTs). (c) Sketch of the waveforms of two type of events. The different ratios of the charge (S2) and the light (S1) signal allow the discrimination between nuclear recoils from WIMPs and neutrons and electronic recoils from γ- and β-background. (Reproduced with permission of [855, 872].)

nuclei ($\sim \times 10^{-12}$ s), resulting in different wave forms. Therefore, by just looking at the wave shape it is possible to differentiate nuclear recoil from the electron recoil.

(2) **Two-phase Xe detector**: Under a strong electric field, not all the ionized Xes recombine, but some drift and exit from the liquid surface. They are accelerated in the gas Xe inducing electroluminescence (proportional scintillation) (signal S2). It is delayed and more intense relative to S1. As the excited Xe by heavy nuclei loses

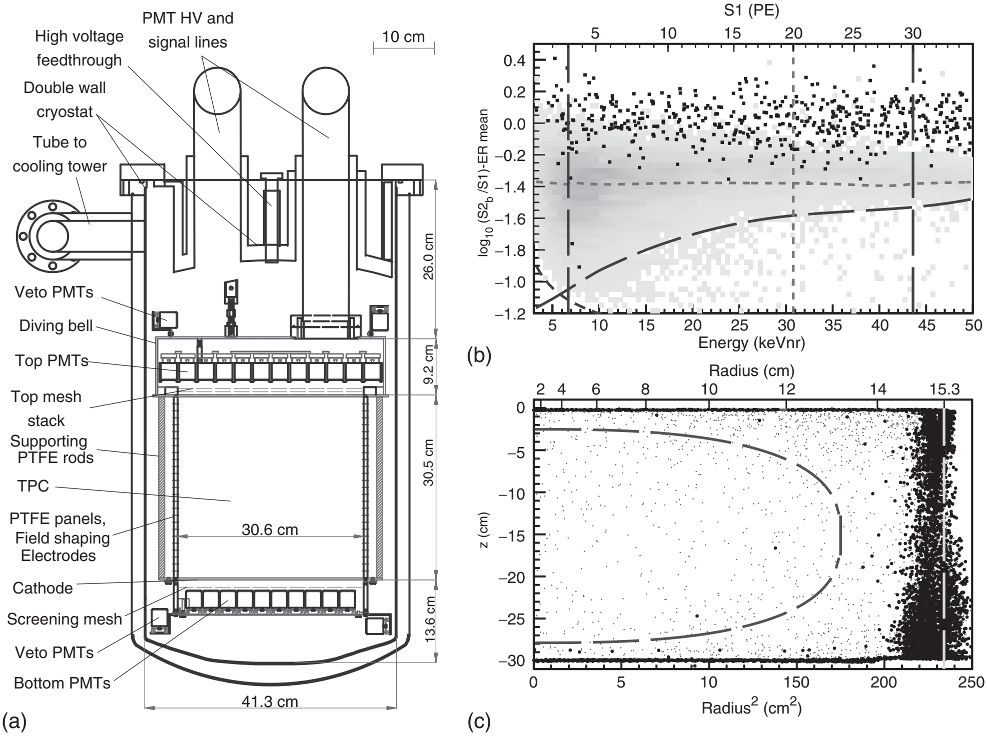


Figure 10.23 (a) Drawing of the XENON100 dark matter detector. The inner TPC contains 62 kg of liquid xenon as target and is surrounded on all sides by an active liquid xenon veto of 99 kg. The diving bell assembly allows keeping the liquid–gas interface at a precise level and enables the filling of LXe in the vessel to a height above the bell. (Reproduced with permission of [872].) (b) Event distribution in parameter space log10(S2b/S1), after using all analysis cuts and a 34-kg fiducial volume (black squares). The analysis uses an upper energy threshold of 43.3 keV (dashed-blue) and a lower threshold of 6.6 keV. The histogram in red/gray indicates the nuclear recoil band from the neutron calibration. An additional hard S2b/S1 cut at 99.75% ER rejection (horizontal dotted green line). Two events fall into the benchmark region, with 1 ± 0.2 is expected from background. PE denotes the number of photoelectrons and ER electron recoil. (c) Spatial event distribution inside the TPC using a 6.6–43.3 keV energy window. The 34-kg fiducial volume is indicated by the red dashed line. Gray points are above the 99.75% rejection line, and the black circles fall below. (Reproduced with permission of [855].)

much of its energy by recombination, the signal S2 is much smaller than that by electron recoil. On the other hand, S1 tends to be larger for the heavy nuclei. Thus the ratio S2/S1 is a good discriminator to pick up signals from backgrounds due to electrons.

Among constructed Xe detectors, Zeplin III [873, 874] and XENON10 [872, 875] emphasize low noise by using two-phase Xenon and obtaining two time-shifted signals, while XMASS [876] emphasize the large mass and self-shielding ability of its active filter. XMASS1 (100 kg fiducial volume and 835 kg total) has operated a very limited number of times, but achieved high light yields of 14.7 pe/keVee and lowest energy threshold of 0.3 keV for the whole volume analysis [877].

As the XENON100 is currently leading the DM search contest, we will describe its detection mechanism. A drawing of XENON100 detector is given in Figure 10.23a. Their event distributions on $E_R - S2/S1$ $\big($recoil energy vs log(S2/S1)$\big)$ and spatial event distribution in the detector are show in Figure 10.23b,c. Two events are in the signal region, with expected backgrounds 1 ± 0.2. Interpreting them as no signals, the upper limit on the DM-nucleon cross section as a function of the DM mass m_χ is given in Figure 10.24.

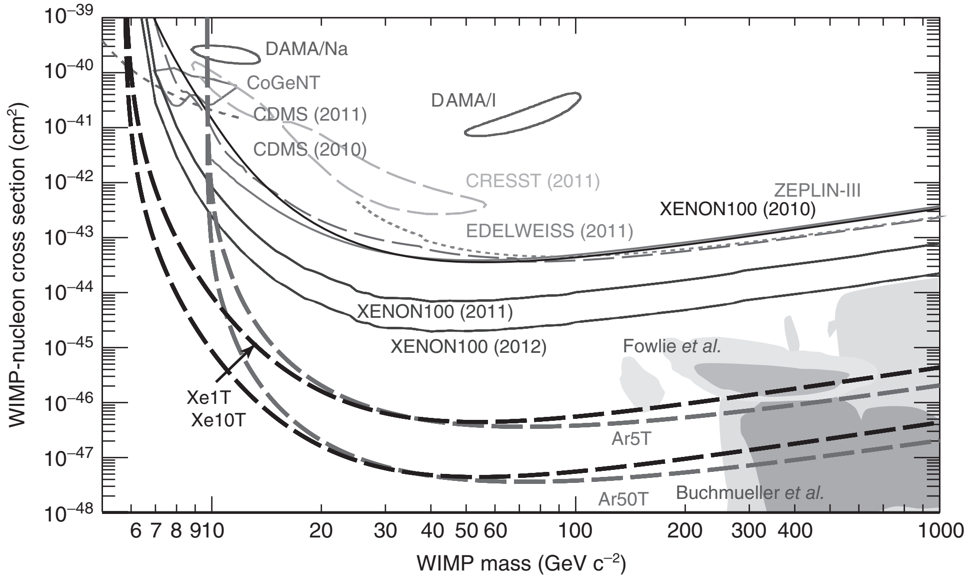


Figure 10.24 Some recent experimental limits on the spin-independent DM-nucleon cross sections and expected future sensitivity for Xe and Ar detectors with fiducial mass of $(m_{\mathrm{Xe}}, m_{\mathrm{Ar}}) = (1, 5)$, $(10, 50)$ tons, plotted assuming future nuclear recoil energy thresholds ~8 keV for Xe and ~20 keV for Ar. Curves show 90% confidence limits for WIMP-nucleus cross section (10^{-46} cm$^2 = 10^{-10}$ pb) for a life time of 1 year. The shaded regions ($1\sigma/2\sigma$) are preferred by supersymmetric (CMSSM, NUHM) models with inputs from recent LHC data [884, 885]. (Reproduced with permission of [855, 867].)

10.5.4
Current Status

At present, the best sensitivity has been obtained by detectors with a combination of ionization loss and bolometer EDELWEISS [878, 879], CDMS [871, 880]) and Xe detectors ZEPLINIII [874] ans XENON100 [855]. Positive data claimed by DAMA [881] with positive annual modulation effect falls in the excluded region by those groups and was not confirmed.

Finally, by extrapolating the current performance, we plot the possible experimental reach of an ideal 1 ton/10 ton Xe and 5 ton/50 ton Ar as dashed lines in Figure 10.24 [867]. The allowed parameter range of the supersymmetric models are narrowing as a result of LHC data [882, 883]. We show an example of the preferred region as constrained by the supersymmetric (CMSSM, constrained MSSM; NUHM, Non-Universal Higgs Masses) models using recent LHC data which are indicated by the shaded regions [884].

11
Dark Energy

The dark energy is the most mysterious entity whose existence was clarified at the turn of the century. It is the source of the accelerating expansion of the universe or, in simple words, the source of antigravity. Its true identity is not known. The uniform vacuum energy that appears in any quantum field theory (aka the cosmological constant) is the primary suspect, but there is a plethora of models to account for the phenomenon. Research on the dark energy is very much in its infancy. Its existence was clarified only by measuring the velocity of receding supernovae (SNe) and through accumulation of indirect evidences. In this chapter, we discuss their evidence for its existence, its properties, and the experimental methods with the hope of clarifying its properties.

11.1
Dark Energy

11.1.1
Accelerating Universe

In Chapters 8 and 9, we used the dark energy only as a background information to describe cosmological phenomena [886]. However, one should be aware that it is a major component of the present cosmic energy. The whole discussion of the post radiation–matter equality epoch would have been full of contradictions had we neglected its existence.

Historically, the existence of the cosmological constant has been suspect for a while, as suggested by [887] in 1984 to account for the smallness of $\Omega_m \simeq 0.1 - 0.3$ against a strong preference (i.e., inflation model) in favor of $\Omega = 1$. Other hints were the evolution of the large-scale structure and the age controversy (see Eq. (11.2)). The suspicion grew strong in the 1990s [888, 889].

A definite evidence was provided by the observation of faraway type Ia supernovae (SNIa) which showed unambiguously the accelerating expansion of the universe.

Type Ia Supernova The type Ia supernova SNIa acts as a standard candle in the large distance $z \sim O(1)$ because its properties are well known and its absolute brightness is comparable to that of galaxies. It appears when the peripheral matter

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

of expanding red giants accrete on a companion white dwarf. The dwarf is in a degenerate state and the accretion of matter triggers carbon deflagration to explode as a type Ia supernova (see arguments in Section 2.9.1). As there are a variety of SNe (type II, Ib, Ic, etc.), one needs to sample the correct type SNIa to use them as the standard candle. They can be identified by their characteristic light spectrum: the absence of hydrogen lines, the existence of silicon lines at $\lambda \simeq 0.62\ \mu$m, and the characteristic luminosity curve (Figure 11.1). The time variation of the spectrum is well known, and with proper calibration the absolute luminosity can be obtained. Therefore, by identifying SNIa and knowing its red shift, one can obtain the acceleration rate of the past universe. The frequency of SNe explosion is estimated once per 20–100 years in a galaxy and the number of observed explosions was limited in the past. With the advent of the CCD (charge-coupled device, elements used in digital cameras), automated image recording of the sky is now possible.

Detection of SNe can be performed as follows: Photographs of a portion of the sky are taken with some time interval (say 2 weeks). By computer processing, the

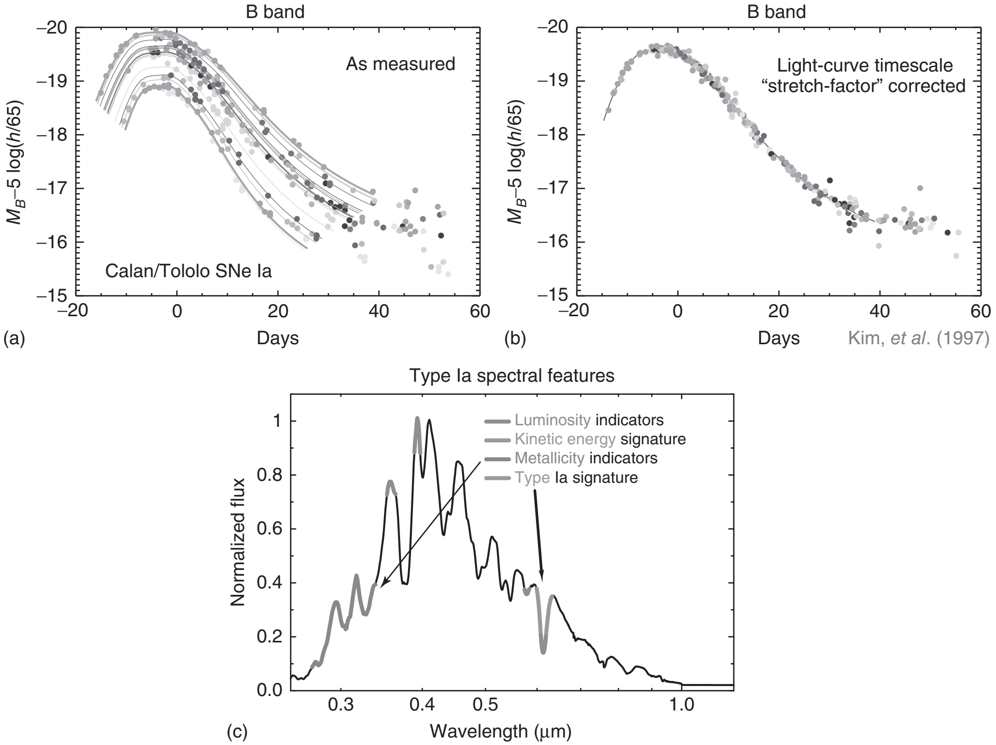


Figure 11.1 (a,b) Light curves as a function of time. The raw spectra show variations but after calibration one obtains a unique absolute luminosity curve which can serve as a standard candle. (c) Typical spectrum of the SN Ia characterized by the absence of the hydrogen line and existence of the Si line. Luminosity and metallicity indicators are also designated (Reproduced with permission of [890, 891].) (Please find a color version of this figure on the color plates.)

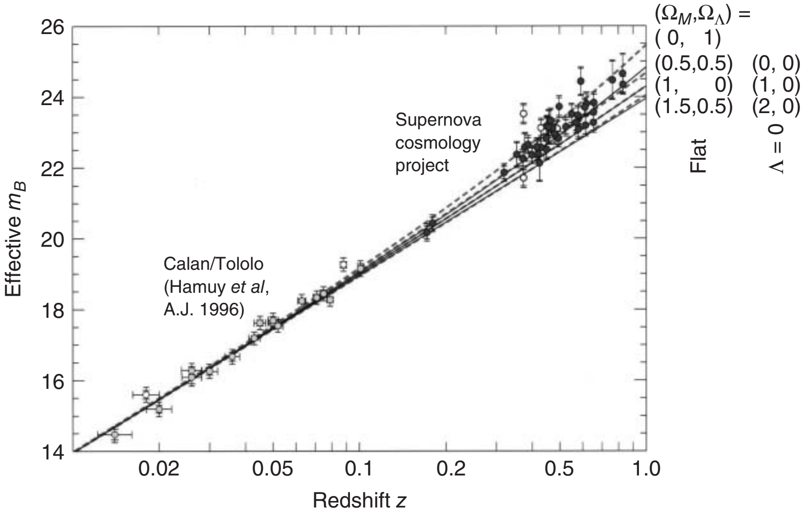


Figure 11.2 Effective apparent luminosity m_B versus the redshift z. Ordinate is the effective brightness m_B. The solid curves are the theoretical $m_B(z)$ for a range of cosmological models with zero cosmologicalconstant: $(\Omega_M, \Omega_\Lambda)$ = (0,0) (top), = (1, 0) (middle), and (2, 0) (bottom). The dashed curves are for a range of flat cosmologicalmodels: $(\Omega_M, \Omega_\Lambda = (0, 1)$ at the top, $(\Omega_M, \Omega_\Lambda) = (0.5, 0.5)$ second from the top, (1, 0) third from the top, and (1.5, 0.5) at the bottom (Reproduced with permission of [894–896].)

image of the first picture is subtracted from the second and the residual bright spot is identified as a supernova. Thus equipped with modern technology, two teams, the "High-z Supernova Search Team" [892, 893] and the "Supernova Cosmology Project" [894] found evidences for the accelerating universe.[1] Figure 11.2 illustrates that the Hubble diagram (distance vs red shift relation) is now extended beyond $z = 1$ and shows the power of SNIa as a standard candle to determine the distances of objects deep in the universe.

Figure 11.3 shows the distance modulus μ (apparent luminosity) versus the red shift z of SNe discovered by ground-based detectors and HST (Hubble Space Telescope). The data fitted with the standard ΛCDM (see Section 11.1.3) model are shown in the inset.

The data clearly indicates transition from deceleration (points below the horizontal line) to acceleration (points above the horizontal line) at $z \simeq 1.3$. Deviation from the old standard model ($\Omega_M = 1$, $\Omega_\Lambda = 0$) is also evident.

11.1.2
Cosmic Age

One drastic outcome of the accelerating expansion was the re-evaluation of the cosmic age. Let us calculate the cosmic age in the model without the cosmological

---

[1] For their achievements, the 2011 Nobel Prize in Physics was awarded to S. Perlmutter, B.P. Schmidt, and A.G. Riess.

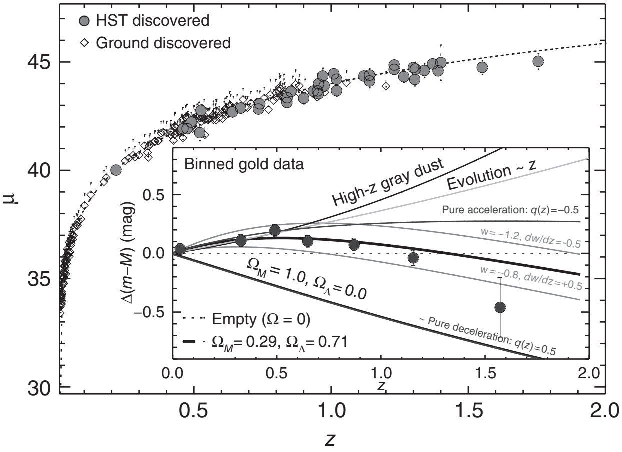


Figure 11.3 SNe Ia from ground-based discoveries are shown as diamonds, and HST-discovered SNe Ia are shown as filled symbols. Overplotted is the best fit for a flat cosmology: $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$. Inset: Residual Hubble diagram and models after subtracting the empty universe model. The distance modulus μ is defined by Eq. (9.8) (Reproduced with permission of [897, 898].)

constant. The present age t_0 of the universe is given by

$$t_0 = \int_0^{t_0} dt = \int_0^{R_0} \frac{dR}{H(t)R} = \int_0^{\infty} \frac{dz}{H(z)(1+z)} \tag{11.1}$$

As the radiation era is short [$\sim$50, 000 years; see Eq. (8.135)] compared to the age of the universe, we can calculate Eq. (11.1) assuming the matter dominated and flat ($k = 0 \rightarrow \Omega_0 = 1$) universe which was a standard assumption before the discovery of the dark energy. The Friedmann equation gives $H(z) = H_0(1+z)^3$ and we obtain

$$t_0 = \frac{2}{3}\frac{1}{H_0} \simeq 9\,\mathrm{Gyr}\,(9\times = 10^9\mathrm{yr}) \tag{11.2}$$

This created an age crisis for the cosmology because measured ages of some observables described below exceeded this number. Obviously, off springs cannot be older than their mother.

Analysis of Radioisotopes One method to determine the age of the universe is to analyze components of long-life radio isotopes (uranium U, thallium Th, and lead Pb) contained in meteors.

$$\begin{aligned} {}^{232}\mathrm{Th}: &\quad t_{1/2} = 1.405 \times 10^{10}\ \mathrm{yrs} \\ {}^{238}\mathrm{U}: &\quad t_{1/2} = 4.468 \times 10^{9}\ \mathrm{yrs} \\ {}^{235}\mathrm{Th}: &\quad t_{1/2} = 7.038 \times 10^{8}\ \mathrm{yrs} \\ {}^{244}\mathrm{Pb}: &\quad t_{1/2} = 8.26\ \ \times 10^{7}\ \mathrm{yrs} \end{aligned} \tag{11.3}$$

These heavy elements were produced in r-processes (r for rapid) at the time of supernovae explosions, and were scattered in the universe. Their remnants observed today have gone through partial decays emitting alpha and beta rays. Incorporating the production/decay rate in the stellar evolution theories, we can estimate the age of galaxies by measuring components ratio of several heavy elements [899].

$$11.8 \text{ Gyrs} \le t_G \le 19.8 \text{ Gyrs} \quad (\text{Gyrs} = 10^9 \text{ yrs}), \qquad \text{average} \sim 14 \text{ Gyrs} \quad (11.4)$$

Globular Clusters Another deduction of the age comes from stellar evolution theory. Stars in a globular cluster are considered to have been born at the same time. The evolutionary stage of the stars can be well represented by the Hertzsprung-Russel diagram. It is a scatter plot on the surface temperature-luminosity plane [See Figure 11.4a]. Traditionally the surface temperature was specified by the color index B-V. When the hydrogen is burning at the core of the stars (main sequence), they align themselves on the Hertzsprung-Russel diagram in a line going upwards from bottom right to upper left.

Although they are born at the same time, massive stars are bluer (i.e., at higher surface temperature) and have shorter lives. This is because despite its larger fuel stock ($\propto M$) its consumption rate $\propto M^{3-5}$ is even faster due to higher core pressure/temperature to counter balance the gravity. As the cluster ages, the stars which were positioned at the left upper corner in the beginning enter helium burning stage faster and peel off to a branch going up to the right becoming red

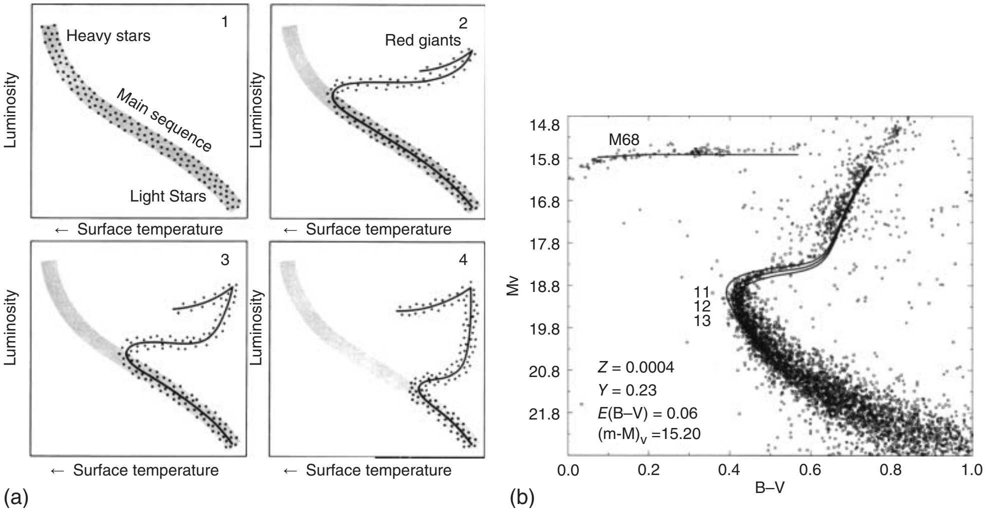


Figure 11.4 (a) Hertzsprung-Russel chart: Distribution of stars classified by their color (surface temperature) and absolute luminosity. Stars in a globular clusters were considered as born at the same time. The four charts show evolving distributions from top left at the birth of globular star to bottom right when almost all the stars grew old. The kink at the branch from the main sequence to the red giants is a good indicator of the age of the globular cluster. (b) Hertzsprung-Russell diagram of globular clusters (GC) M68. Theoretical lines with the age given in Gyrs are overlayed. See also Figure 7.9 for more details of the stellar evolution. (Reproduced with permission of [900].)

giants. The kink at the branch from the main sequence to the red giants is a good indicator of the age of the globular cluster.

Figure 11.4b shows an example of such distributions of stars in a globular cluster [900]. The horizontal branch on top is an intermediate stage from red giants to asymptotic giant branch (AGB). The line is separate because its transition period is short which results in sparse population. The ages deduced from observations of the globular clusters is [901]

$$t_G = 12.9 \pm 2.9\,\text{Gyrs} \tag{11.5}$$

One sees that the cosmic ages determined by both isotopes and globular clusters are older than one predicted from the cosmic expansion rate using the matter dominant flat universe model.

The age crisis can be avoided by including the dark energy *i.e.* cosmological constant term. As the expansion rate is increasing at present time, the scale evolution curve $R(t)$ as a function of time has to be concave upward as shown in Figure 11.5 (see also Figure G2). Obviously, inclusion of the cosmological constant prolongs the age.

Using Eq. (8.68), Eq. (11.1) can be integrated analytically for $k = 0$.

$$t = \frac{1}{H_0} \int_z^\infty \frac{dx}{(1+z)[\Omega_\Lambda + \Omega_{m0}(1+z)^3]^{1/2}} \tag{11.6}$$

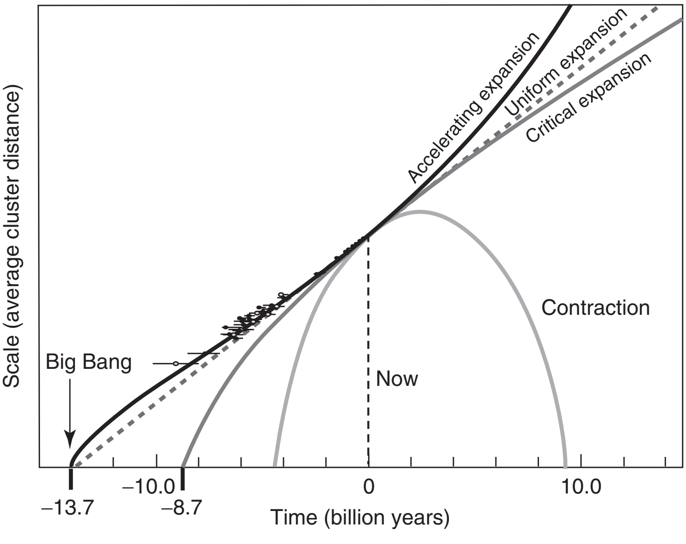


Figure 11.5 History of cosmic expansion, as measured by the high-redshift supernovae (black data points). Given energy compositions, time evolution curves can be calculated from the Friedmann equation, assuming flat cosmic geometry. Without dark energy, acceleration is always decelerating and the curve is convex upward. With dark energy, the expansion is decelerating in the beginning, turns to acceleration later. [1, 890]

By changing the integration variable to $y = \sqrt{\frac{\Omega_\Lambda}{\Omega_{m0}}}(1+z)^{-3/2}$

$$t = \frac{2}{3}\frac{1}{H_0\sqrt{\Omega_\Lambda}}\int_0^{\sqrt{\frac{\Omega_\Lambda}{\Omega_{m0}}}(1+z)^{-3/2}} \frac{dy}{\sqrt{1+y^2}}$$

$$= \frac{2}{3}\frac{1}{H_0\sqrt{\Omega_\Lambda}}\left[\log(y+\sqrt{y^2+1})\right]_0^{\sqrt{\frac{\Omega_\Lambda}{\Omega_{m0}}}(1+z)^{-3/2}}$$

$$= \frac{2}{3}\frac{1}{H_0\sqrt{\Omega_\Lambda}}\log\left[\sqrt{\frac{\Omega_\Lambda}{\Omega_{m0}}}\left\{(1+z)^{-3/2}+\sqrt{(1+z)^{-3}+\frac{\Omega_{m0}}{\Omega_\Lambda}}\right\}\right] \tag{11.7}$$

By setting $z = 0$, we obtain t_0.

$$t_0 = \frac{2}{3}\frac{1}{H_0\sqrt{\Omega_\Lambda}}\log\left[\frac{1+\sqrt{\Omega_\Lambda}}{\sqrt{1-\Omega_\Lambda}}\right] \overset{\Omega_\Lambda=0.74}{=\!=\!=} 1.04\,H_0^{-1} = 13.7\,\text{Gyr}^{\,2)} \tag{11.8}$$

This is a number that does not contradict with the observations. We remark that the cosmic age turned out to be very close to the present Hubble time.

11.1.3
ΛCDM Model

The clear evidence of the accelerating universe discovered by the SNIa observations changed the old standard model ($\Omega_m = 1$, $\Omega_\Lambda = 0$) drastically. The model incorporating a finite cosmological constant and the cold dark matter is referred to as the ΛCDM model. The present standard model is a ΛCDM assuming a flat universe with $\Omega_\Lambda \simeq 0.70$, and $\Omega_m \simeq 0.30$, $\Omega_m + \Omega_\Lambda = 1$.[3] Soon after the SNIa data, very precise data on CMB (cosmic microwave background) spectrum was published by WMAP (Wilkinson microwave anisotropy probe) in 2003 [705, 904], which fitted the ΛCDM model quite well. All the controversies up to that time, namely a theoretical cosmic age too short to explain the age of globular clusters, the matter composition showing much smaller value ($\Omega_m \sim 0.3 < 1$) than expected, growth rate of large-scale structure being much faster than predicted, etc., were solved. It was a revolution in cosmology, similar to the establishment of the Standard Model of particle physics in 1970s. The ΛCDM model incorporating precise values of cosmic parameters as determined by the WMAP is now the standard of the cosmological model and is referred to as the *concordance model*.[4] Figure 11.6a shows a constrained region on $\Omega_m - \Omega_\Lambda$ plane by the SNe, WMAP, and BAO (baryon acoustic oscillation: a trace of CMB imprinted in the large scale structure) data. The significance of various cosmic models depending on where thery are located on $\Omega_m - \Omega_\Lambda$ plane is illustrated in Figure 11.6b. Its details is discussed in Appendix G.

---

[2] Gyr=10^9 years.

[3] As the cosmological constant is only one candidate among many to cause the accelerating universe, a more generic name "dark energy" was used by Turner [902, 903] and it prevailed.

[4] Recently more accurate data obtained by the Planck have become available [123].

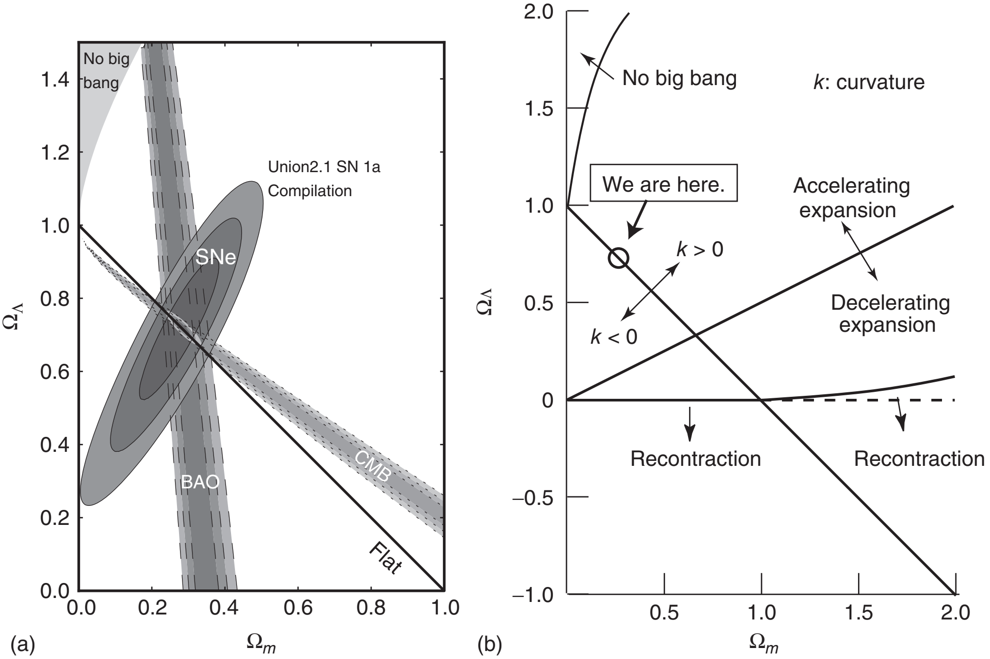


Figure 11.6 (a) Constraints on Ω_m and Ω_Λ in the concordance model using baryon acoustic oscillations (BAO), cosmic microwave background (CMB), and supernovae (SNe) measurements. Curves indicate 68.3%, 95.4%, and 99.7% confidence regions. Cosmological constant dark energy ($w = -1$) has been assumed. (Reproduced with permission of [905, 906].) The datasets are WMAP [741, 781], SNIa [892–895, 897], and BAO [907]. (b) The position of the concordance model is shown on the $\Omega_m - \Omega_\Lambda$ plane where many cosmic models are separated. k denotes the curvature of the universe. For an explanation of the "no big bang" at the upper left, see Appendix G.

11.2
Cosmological Constant

It is a well-known fact that Einstein introduced the cosmological constant to construct a stationary universe and later abandoned it after the discovery of the expanding universe. As we discussed in Chapter 8, it can be interpreted as a homogeneous vacuum energy with negative pressure given by[5]

$$\rho_{VAC} = \frac{\Lambda}{8\pi G_N}, \quad P_{vac} = -\rho_{VAC}. \tag{11.9}$$

The cosmological constant is the simplest explanation for incorporating the accelerating universe, and phenomenologically there is no reason to deny it. However, it does not explain the dynamical origin of the repulsive force. Furthermore, it suffers from a couple of conceptual difficulties, namely the "cosmological constant problem" and the "coincidence problem".

---

[5] The cosmological constant as a mathematical entity can be positive or negative to represent a repulsive or an attractive force.

The cosmological constant problem: The vacuum energy has a natural and logical origin as the quantum fluctuation of fields. The energy scale for the quantum origin of the Big Bang is the Planck energy $E_{\rm Pl} = G_N^{-1/2} \sim 10^{19}$ GeV. As the energy density has dimension $[E^4]$ in natural units, the most natural vacuum energy density that can be inferred theoretically is

$$\rho_{vac}\,(\text{predicted}) \simeq M_{\rm Pl}^4 \simeq 10^{76}\,\text{GeV}^4. \tag{11.10}$$

On the other hand, the dark energy density as we observe today is

$$\rho_{vac}(\text{observed}) \simeq 0.70\rho_c = 0.70 \times 1.88 \times 10^{-29} h^2\ \text{g/cm}^{-3} \simeq 10^{-47}\,\text{GeV}^4. \tag{11.11}$$

As the cosmological constant is a constant by definition, it should have existed from the very beginning. There exists a mismatch of 120 digits between the prediction and the observation! A conventional logic to overcome such problem is to require a symmetry that makes the vacuum energy to vanish. Indeed, the boson fields produce $E_B = \frac{1}{2}\sum_{\boldsymbol{k}} g_B \sqrt{\boldsymbol{k}^2 + m_B^2}$, while the fermion fields produce $E_F = -\frac{1}{2}\sum_{\boldsymbol{p}} g_F \sqrt{\boldsymbol{p}^2 + m_F^2}$. The sum vanishes if there is exact symmetry between bosons and fermions, that is, the supersymmetry. If we did not discover the small dark energy, we would have been happy. But a finite residual energy as tiny as 10^{-120} is unthinkable. From arguments in particle physics, the supersymmetry is expected to be broken at the energy scale $\sim$ 1 TeV. Then the predicted energy imbalance is $\sim(1000\ \text{GeV})^4 \sim 10^{12}\ \text{GeV}^4$, which is a large reduction by a factor 10^{60}, but still $\times 10^{60}$ larger than the observed value.

The coincidence problem: Another problem is the coincidence problem. The matter energy varies like $(1+z)^{-3}$ as time goes, the radiation energy $\sim(1+z)^{-4}$, and the curvature term $\sim(1+z)^{-2}$, while the vacuum energy stays constant. The most dominant component at a particular epoch is usually larger than others by many orders of magnitude, as illustrated by Figure 11.7.

The radiation energy $\Omega_{\rm rad} \propto R^{-4}$ was the most important energy contributor in the early universe. The Ω_Λ occupied only a fraction of 10^{-100} at the GUT era but it is now $\sim$0.74 and is comparable with $\Omega_{m0} \sim 0.26$. Notice, however, that it has become dominant only recently at $z \sim 0.8$ which is, figuratively speaking, a split second before the present time in the long history of the universe.

Then, in the long history of the universe $0 < z < 10^{30}$,[6] why the dark energy and matter energy should be of the same order at the particular time now ($0 \lesssim z \lesssim 1$)? Why has Nature provided the vacuum energy as small as 120 orders of magnitude

---

[6] We are counting the time logarithmically. $z = 10^{30}$ corresponds to the cosmic temperature of $\sim 10^{19}$ GeV. If we start counting from the nucleosynthesis ($T \sim 0.1$ MeV), which happened at $t = 1$ s, $z \sim 10^9$.

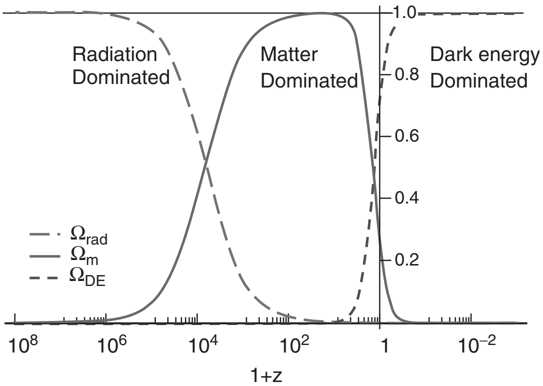


Figure 11.7 Energy components as a function of z. The present time is at $1 + z = 1$.

in the very beginning, hide it for a long time, and made it appear at the particular time now? It is too much of a coincidence to be true.

Anthropic Principle: Some people rely on the anthropic principle to answer the question [908]. It depends on the assumption that our universe is just one member of an ensemble.

At the root of physics, there are several fundamental constants such as the light velocity c, the Planck constant $\hbar$, and coupling constants of the fundamental interactions. We do not know why they take the values they have. If their values are different, physical phenomena would be quite different even if the laws of physics were the same. As there is no criterion to constrain their values, we may consider that universes with different constants could exist, too. The inflation offers such a scenario because the phase transition to new vacua at the end of the inflation era is like bubbling and could have created a vast number of universes under slightly different conditions.

There is no rule to constrain the orbital radii of planets in the Newtonian law. The Earth is special among the many planets and happened to have the right environment to nurture life, eventually producing human beings. Stated differently, the radius of the earth had to take its value to allow the existence of human beings. This would not be a satisfactory logic if the Earth were the only planet in the universe. But with eight more planets in our solar system and the vast numbers of planets in the universe, the logic makes sense. Likewise, we may consider our universe also special in the sense that it happens to have the right physical constants to produce the phenomena we are observing. All the values of the fundamental constants can be explained simply by the fact that we live in this universe.

The anthropic logic can provide an answer to the coincidence problem as well as to the cosmological problem. The fact that the value of the cosmological constant is very close to the present cosmic matter density can be explained as follows. If its value were large, there would be no era of matter dominance or at least it would shorten its time span preventing formation of galaxies, hence, no planets and no living creatures. The argument can be made more precise by considering how long it takes to develop an intelligent life to ask the questions we are asking now. Calculations based on those logics produce values surprisingly close to the observed ones.

The anthropic principle seems capable of finding an answer to any question. But it may not predict anything in which, we believe, physics justification lies.

Models of Acceleration Mechanism Understanding the origin of cosmic acceleration is a challenge for particle physicists as well as cosmologists. The original cosmological constant as conceived by Einstein is one candidate. However, it is not necessarily the only one. A smooth component of fluids with large negative pressure can provide the repulsive gravity and can lead to the observed accelerated expansion within the context of general relativity. There is plethora of ideas of what the dark energy might be, ranging from the quantum vacuum to a new ultralight scalar field. Some have proposed to modify the general relativity or to introduce extra spatial dimensions. Here, we will limit our discussions to the cosmological constant and a new scalar field generically dubbed as the quintessence.

11.3
Quintessence model

The quintessence[7] model offers a way out of the coincidence problem. It is a time-varying scalar field having a very long range force [909, 910]. It has a tracker behavior in the sense that the quintessence field has a energy density that closely tracks (but is less than) the radiation density until matter–radiation equality, which triggers quintessence to start having characteristics similar to the dark energy, eventually dominating the universe.

Let us start from the Lagrangian of the scalar field ϕ, which takes into account of general relativity, that is, expanding universe.

$$S = \int d^4x \sqrt{-\det g_{\mu\nu}} \left[\frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi) \right] \tag{11.12}$$

where $g_{\mu\nu}$ is the Robertson–Walker metric tensor $g_{\mu\nu}$ given by Eq. (8.2). The energy and the pressure (energy–momentum tensor) can be calculated as

$$\rho_\phi = \frac{1}{2} \dot{\phi}^2 + V(\phi), \tag{11.13a}$$

$$P_\phi = \frac{1}{2} \dot{\phi}^2 - V(\phi) \tag{11.13b}$$

---

[7] Quintessence means "the fifth element" or the essence of a thing in Latin. The four elements are air, earth, fire, and water in ancient Greek science, or baryons, leptons, photons, and dark matter in modern terminology.

where $\dot{\phi} = d\phi/dt$, and the field is assumed to depend only on time, not on position. One can see that the equation of state parameter is given by

$$w_\phi = \frac{P_\phi}{\rho_\phi} = \frac{\frac{1}{2}\dot{\phi}^2 - V(\phi)}{\frac{1}{2}\dot{\phi}^2 + V(\phi)} \quad \rightarrow \quad \begin{cases} -1 & \dot{\phi}^2 \ll V \\ +1 & \dot{\phi}^2 \gg V. \end{cases} \tag{11.14}$$

If the potential energy is larger than the kinetic energy, the pressure can be negative. Equation (11.14) also ensures that $w_\phi > -1$. Suppose the field had initially a much larger kinetic energy but later it became smaller than the potential energy; then the field will acquire a negative pressure at later stage. The evolution of the ϕ field density will be accelerated and will eventually become the dominant component of the cosmic energy density. Therefore, a scalar field with a potential having a steep slope in the beginning but slowing down in later stage is a candidate for the dark energy (see Figure 11.8).

Slow Rolling Let us see what conditions the quintessence field has to have. The field equation can be obtained from the action Eq. (11.12) [or alternatively combining Eqs. (11.13) with the continuity equation Eq. (11.47)].

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0 \tag{11.15}$$

where $V' = \partial V/\partial\phi$. The requirement $\dot{\phi}^2 \ll V$ and $\ddot{\phi} \ll V'$, where the latter follows from the former, are the same conditions as the slow rolling we required for the inflation model [see Eq. (8.152)].

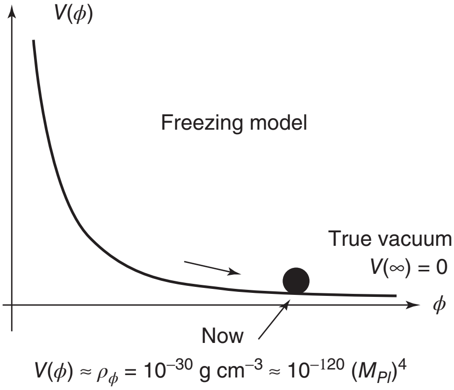

$V(\phi)$
Freezing model
True vacuum
$V(\infty) = 0$
ϕ
Now
$V(\phi) \approx \rho_\phi = 10^{-30}$ g cm$^{-3} \approx 10^{-120}\,(M_{Pl})^4$

Figure 11.8 Typical quintessence potential for the freezing (tracking) model. The universe is rolling down the potential. When it reaches a point where $\dot{\phi}^2/2 \ll V(\phi)$, the pressure becomes negative and the universe will enter the accelerating expansion mode. It then moves slowly, almost freezing. The present value of the potential is $V \sim \rho_\phi \simeq 10^{-30}\ \mathrm{g\,cm^{-3}} \sim 10^{-120} M_{\mathrm{Planck}}^4$. Eventually, it reaches the true vacuum state, which is assumed to have zero energy. But a finite asymptotic value either positive or negative is not excluded.

The parameter w_ϕ can be expressed as

$$1 + w_\phi = \frac{\dot{\phi}^2}{\rho_\phi} = \frac{\dot{\phi}^2 V'^2}{V'^2 \rho_\phi} = \frac{\dot{\phi}^2 V'^2}{(\ddot{\phi} + 3H\dot{\phi})^2 \rho_\phi} = \frac{V'^2}{9H^2(1+\xi_\phi)^2 \rho_\phi} \tag{11.16}$$

where $\xi_\phi \equiv \ddot{\phi}/(3H\dot{\phi})$. As, in the slow rolling limit, $|\xi_\phi| \ll 1$ and $\rho_\phi \simeq V$, we have

$$1 + w_\phi \simeq \frac{V'^2}{9H^2 \rho_\phi} = \frac{1}{3}\frac{V'^2}{8\pi G_N \rho_{\text{tot}} \rho_\phi} \simeq \frac{1}{3}\frac{V'^2}{8\pi G_N V^2} \equiv \frac{1}{3}\lambda^2 \tag{11.17}$$

where the Friedmann equation Eq. (8.10) was used in going to the third equality. The quintessence dominance was assumed in obtaining the penultimate equality because it is the condition we require eventually. Notice that $\lambda^2/2$ coincides with the slow rolling parameter ε which we defined in the discussion of the inflation [Eq. (8.154)]. In order to have accelerating expansion, we need $\rho + 3P < 0$ [see Eq. (8.16)] or $w < -1/3$, which, in turn, requires $\lambda^2 < 2$. Phenomenologically, $w \sim -1$ is realized, and therefore $\lambda^2 \ll 1$.

The parameter λ characterizes the slope of the field potential and its time evolution rate, or, equivalently, its derivative with respect to $\ln R$ satisfies the condition

$$\frac{d\lambda}{d\ln R} = R\frac{dt}{dR}\frac{d\phi}{dt}\frac{d\lambda}{d\phi} = \frac{\dot{\phi}}{H}\frac{d\lambda}{d\phi} = -\lambda^2(\Gamma - 1)\frac{\kappa\dot{\phi}}{H} \tag{11.18a}$$

$$\Gamma \equiv \frac{VV''}{V'^2}, \qquad \kappa \equiv \sqrt{8\pi G_N}. \tag{11.18b}$$

If $\Gamma = 1$, then $\lambda =$ constant and the potential takes an exponential form $V(\phi) = V_0 e^{\pm\kappa\lambda\phi}$. In this case, $w_\phi = w_B$, where suffix B denotes the dominant component (radiation or matter) and the energy density of the ϕ field is proportional to it [see Eq. (11.22)]. It is referred to as the *scaling solution* [911]. This is the case for accelerating cosmic expansion with power law ($R \propto t^\gamma$ ($\gamma > 1$)). (See Problem 11.1.)

Problem 11.1 *(1) Apply Eq. (11.13) to the Friedmann equation Eq. (8.10) and the equation of cosmic acceleration Eq. (8.16) and show that*

$$\begin{aligned} \dot{\phi}^2 &= -\frac{1}{4\pi G_N}\dot{H} \\ V(\phi) &= \frac{3}{8\pi G_N}\left(H^2 + \frac{\dot{H}}{3}\right). \end{aligned} \tag{11.19}$$

(2) For $R \propto t^\gamma$, $\gamma > 1$, solve Eq. (11.19) to obtain

$$\begin{aligned} \phi &= \left(\frac{\gamma}{4\pi G_N}\right)^{\frac{1}{2}} \ln t + \text{const.}, \\ V(\phi) &= \frac{3}{8\pi G_N}\frac{\gamma(\gamma - \frac{1}{3})}{t^2} = V_0 \exp\left[-\sqrt{\frac{2}{\gamma}}\,\kappa\phi\right]. \end{aligned} \tag{11.20}$$

If Γ is larger than 1, and if it stays nearly a constant for a span of many Hubble times, namely if

$$\left|\frac{d(\Gamma-1)}{dt}\right| \ll H|\Gamma-1| \tag{11.21}$$

which is referred to as the *tracking condition*, then the value of λ will eventually become small enough to realize dominance of the ϕ field.

We also have to make sure that the quintessence is a subdominant component in the early stage.[8] The condition is satisfied if $w_\phi < w_B$. As the energy density evolves as $\rho \sim R^{-3(1+w)}$ [see Eq. (8.53)], the quintessence evolves more slowly than the dominant radiation energy and ensures that it is a subdominant component in the early stage. It also ensures that it becomes the dominant component eventually because w_ϕ goes negative when the universe is matter-dominated.

In fact, for the fields satisfying the tracking condition, it is shown that the following equality holds [912–914]:

$$w_\phi \simeq \frac{w_B - 2(\Gamma-1)}{1+2(\Gamma-1)}. \tag{11.22}$$

Therefore, for $\Gamma > 1$, we have $w_\phi < w_B$ as required.

Proof of Eq. (11.22). We define a parameter x by

$$x \equiv \frac{1+w_\phi}{1-w_\phi} = \frac{\dot{\phi}^2}{2V}. \tag{11.23}$$

x is a small number and if we require

$$\gamma \equiv \frac{d\ln x}{d\ln R} = \frac{1}{Hx}\frac{dx}{dt} \ll 1, \qquad \frac{d\gamma}{d\ln R} \ll 1 \tag{11.24}$$

it changes slowly relative to the Hubble time. Considering the similarity of the variable x with λ, the conditions are the same requirements as λ is small and stays nearly constant [see Eq. (11.17)]. That is, we are requiring the tracker solution.

For the flat universe, the Friedmann equation Eq. (8.10) and Eq. (8.19) mean

$$H^2 = \frac{8\pi G_N}{3}\left(\rho_\phi + \rho_B\right), \qquad \dot{\rho} = -3H(\rho + P). \tag{11.25}$$

Then we have

$$\frac{\dot{H}}{H^2} = -\frac{3}{2}\Omega_\phi(1+w_\phi) - \frac{3}{2}\Omega_B(1+w_B), \tag{11.26a}$$

$$\frac{\kappa\dot{\phi}}{H} = \sqrt{\frac{\kappa^2}{H^2}\dot{\phi}^2} = \sqrt{3\frac{\kappa^2}{3}\frac{\rho_\phi}{H^2}(1+w_\phi)} = \sqrt{3\Omega_\phi(1+w_\phi)}. \tag{11.26b}$$

---

[8] Phenomenologically, the most severe constraint on Ω_ϕ comes from the nucleosynthesis. If the extra scalar energy component is substantial, the cosmic expansion rate changes and modifies the abundance of light elements.

Next we will see how the parameter γ is related to $\lambda = -V'/(\kappa V)$.

$$\gamma \equiv \frac{d\ln x}{d\ln R} = \frac{1}{Hx}\frac{dx}{dt} = \frac{1}{Hx}\left(\frac{\dot\phi\ddot\phi}{V} - \frac{\dot\phi^3 V'}{2V^2}\right) = \frac{1}{Hx}\frac{\dot\phi}{V}\left(\ddot\phi - \frac{\dot\phi^2 V'}{2V}\right)$$

$$\overset{(11.15)}{=} \frac{1}{Hx}\frac{\dot\phi}{V}\left(-3H\dot\phi - V' - \frac{\dot\phi^2 V'}{2V}\right) \overset{(11.23)}{=} -6 - \frac{2}{H\dot\phi}\frac{V'}{V}\left(\frac{1}{2}\dot\phi^2 + V\right)$$

$$= -6 + 2\frac{\kappa\rho_\phi}{H\dot\phi}\left(-\frac{V'}{\kappa V}\right) \overset{(11.17)}{=} -6 + 2\frac{\kappa\rho_\phi}{H\dot\phi}\lambda \tag{11.27}$$

$$\therefore \quad \lambda = \frac{3H\dot\phi}{\kappa\rho_\phi}\left(1 + \frac{\gamma}{6}\right) = \sqrt{\frac{3\dot\phi^2}{\rho_\phi}\frac{3H^2}{\kappa^2\rho_\phi}}\left(1 + \frac{\gamma}{6}\right) = \sqrt{\frac{3(1 + w_\phi)}{\Omega_\phi}}\left(1 + \frac{\gamma}{6}\right). \tag{11.28}$$

For the tracker, we have $\gamma \ll 1$ and hence, the tracker solution is chracterized by

$$\Omega_\phi \simeq \frac{3(1 + w_\phi)}{\lambda^2} \tag{11.29}$$

Equation (11.28) is an alternative expression for λ. We differentiate λ with respect to $\ln R$ and compare it with (11.18) .

$$\frac{d\lambda}{d\ln R} = \frac{d}{d\ln R}\left[\sqrt{\frac{3(1 + w_\phi)}{\Omega_\phi}}\left(1 + \frac{\gamma}{6}\right)\right] = \lambda\left[\frac{1}{2}\frac{w'_\phi}{1 + w_\phi} - \frac{1}{2}\frac{\Omega'_\phi}{\Omega_\phi} + \frac{\gamma'}{6 + \gamma}\right] \tag{11.30}$$

where we used $'$ to denote $d/d\ln R$. Using Eq. (11.23) and $\gamma = d\ln x/d\ln R = x'/x$, the first term in the parenthesis becomes

$$\frac{w'_\phi}{1 + w_\phi} = \frac{\gamma}{1 + x}. \tag{11.31}$$

The second term in the parenthesis can be calculated as follows: Using $\Omega_\phi = (\kappa^2/3)(\rho_\phi/H^2)$ and

$$\begin{aligned} \frac{\Omega'_\phi}{\Omega_\phi} &= \frac{1}{H}\left(\frac{\dot\rho_\phi}{\rho_\phi} - 2\frac{\dot H}{H}\right) \\ &= -\frac{3H(\rho_\phi + P_\phi)}{H\rho_\phi} - 2\left[-\frac{3}{2}\Omega_\phi(1 + w_\phi) - \frac{3}{2}\Omega_B(1 + w_B)\right] \\ &= 3(1 - \Omega_\phi)(w_B - w_\phi). \end{aligned} \tag{11.32}$$

Substituting Eq. (11.31) and (11.32) in Eq. (11.30) and neglecting small terms containing γ and γ', we have an alternative expression for the equation of λ.

$$\frac{d\lambda}{d\ln R} \simeq -\frac{3}{2}\lambda(1 - \Omega_\phi)(w_B - w_\phi) \tag{11.33}$$

Equating Eq. (11.33) with Eq. (11.18) and using Eq. (11.26b) (11.28), we obtain

$$\Gamma - 1 = -\frac{H}{\kappa\dot{\phi}}\frac{1}{\lambda^2}\frac{d\lambda}{d\ln R} = 3\frac{(1-\Omega_\phi)(w_B - w_\phi)}{(1+w_\phi)(6+\gamma)} \simeq \frac{(1-\Omega_\phi)(w_B - w_\phi)}{2(1+w_\phi)}. \tag{11.34}$$

For $\Omega_\phi \ll 1$, Eq. (11.34) gives the desired expression

$$w_\phi \simeq \frac{w_B - 2(\Gamma - 1)}{1 + 2(\Gamma - 1)}. \tag{11.35}$$

All the properties of the field ϕ can be most easily elucidated by examining representative tracking potentials satisfying the tracking conditions [see problem (11.2)]. Typical examples are [909]

$$V_1(\phi) = M^{4+n}\phi^{-n}, \tag{11.36a}$$

$$V_2(\phi) = M^{4+n}\phi^{-n}\exp\left[\frac{\alpha\phi^2}{M_{Pl}^2}\right]. \tag{11.36b}$$

The difference of the two potential is that $V_2(\phi)$ has a minimum at a finite value of ϕ which the field approaches, while $V_1(\phi)$ is rolling forever toward $V(\phi = \infty) = 0$. Note that, if the potential has a constant term V_0, the destiny of the universe is very different depending on its sign. If the potential goes negative, the universe will eventually turn to a Big Crunch.

Problem 11.2 *When a scalar field has the potential expressed as an inverse power of the field* $V(\phi) = M^{4+n}\phi^{-n}$, $n > 0$ *where n need not to be an integer, one can show that it has the properties of the tracker.*

1. *Show* $\Gamma = 1 + 1/n =$ *constant* > 1.
2. *Show that in the radiation-dominated era, it has a solution*

$$\phi = \left(\frac{n(n+2)^2 M^{4+n} t^2}{6+n}\right)^{\frac{1}{(2+n)}} \tag{11.37}$$

whose energy density $\dot{\phi}^2 \sim V \sim t^{-2n/(2+n)}$ *grows less rapidly as time goes back than the dominant radiation component* $\rho_{\text{rad}} \sim (t^{1/2})^{-4}$.
Hint: use $H = 1/2t$ *in Eq. (11.15).*

3. *Show that a small perturbation* $\delta\phi$ *satisfies the equation*

$$\delta\ddot{\phi} + \frac{3}{2t}\delta\dot{\phi} + \frac{(6+n)(1+n)}{(2+n)^2 t^2}\delta\phi = 0. \tag{11.38}$$

(4) Show that Eq. (11.38) has two solutions of the form

$$\delta\phi \propto t^\gamma, \qquad \gamma = -\frac{1}{4} \pm \sqrt{\frac{1}{16} - \frac{(6+n)(1+n)}{(2+n)^2}}. \tag{11.39}$$

As the square root is imaginary for $n > 0$, both solutions decay as $t^{-1/4}$ for increasing t while ϕ itself is increasing. It demonstrates the stability and attractor nature of the field.

4. *Show that, in the matter-dominated era, the solution given by Eq. (11.37) is still valid with a different coefficient*
5. *Show that*

$$w_\phi = \frac{n w_B - 2}{n+2} = w_B - \frac{2(1+w_B)}{n+2} \tag{11.40}$$

where w_B is the equation of state parameter of the dominant energy component. It means that ρ_ϕ decreases less rapidly for large t than ρ_m or $\rho_{\rm rad}$, and ρ_ϕ dominates eventually.

6. *Show that the time (t_{eq}) when $\rho_M \simeq \rho_\phi$ in the matter-dominated era is approximately given by $t_{eq} \sim (1/M_{\rm Pl})\,(M_{\rm Pl}/M)^{(4+n)/2}$.*
Hint: use the Friedmann equation $H^2 \simeq 8\pi G_N \rho_M/3$.
7. *Show that $\phi(t = t_{eq}) \simeq M_{\rm Pl}$.*

The nice feature of the tracker quintessence model is that it is insensitive to the initial conditions. Remember that we did not specify it for the tracking condition. It can vary many orders of magnitude without altering the cosmic history [see Figure 11.9). In particular, it includes the most likely initial condition after inflation, namely thermal equilibrium in which equipartition is realized. In this case, the quintessence field density is nearly equal to any other field. Then counting the number of degrees of freedom, $\Omega_\phi \sim 10^{-3}$.

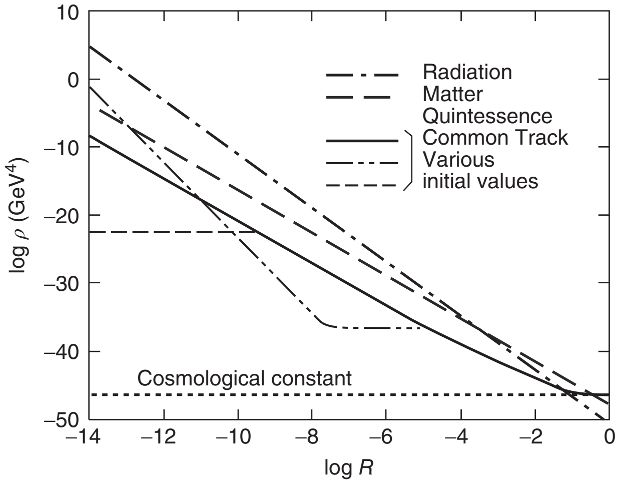


Figure 11.9 Plot of the quintessence energy density for a typical tracker (freezing) model. The solid line represents the evolution that emerges from equipartition at the end of inflation, the dash-dotted line the radiation, and the long dashed line the matter densities. Starting from a variety of initial conditions, the quintessence converges to the common attractor line. [912, 913]

To summarize, the tracker is a scalar field ϕ which rolls down a potential $V(\phi)$ like the one shown in Figure 11.8. It has an attractor-like solution in the sense that fields having a wide range of initial conditions rapidly approach a common evolutionary track. Its energy density remains negligible compared to the radiation density throughout the radiation-dominated epoch, and after entering the matter-dominated epoch, the equation of state parameter turns negative. Once it goes more negative than $-1/3$, the quintessence energy density evolves more rapidly than the matter or radiation and will be the dominant component. After that, the universe enters the accelerating expansion stage. Thus we have a dynamical explanation for the coincidence problem why the dark energy took over in the near past.

Note that as $\lambda = V'/(\kappa V) = M_{\text{Pl}} V'/\sqrt{8\pi} V \sim O(1)$, it means that $|\phi| \sim O(M_{\text{Pl}})$. The field strength at present is of the order of the Planck mass, whereas the potential is minuscule [$V(\phi) \sim \rho_{\text{cr}} \sim 10^{-47}$ GeV^4: see Eq. (11.11)]. The fine-tuning problem is not solved. Although we assumed that the quintessence field depends only on time and not on space, it could be nonuniform spatially. Unlike the vacuum energy, it could exist as a cluster of mass. However, the present mass value is of the order

$$m_\phi \simeq \sqrt{V''}, \simeq \frac{\sqrt{V}}{\phi} \simeq \frac{\sqrt{10^{-47}\,\text{GeV}^4}}{M_{\text{Pl}}} \sim 10^{-61} M_{\text{Pl}} \sim 10^{-32}\ \text{eV} \sim H_0 \qquad (11.41)$$

Its Compton wavelength is of the same order as the size of the whole universe and the spatial uniformity is not lost. It can, however, have a long-reaching force which affects other observables. For instance, if it couples with the photon, its polarization would be affected. Thus in order not to contradict the observations, its interaction with the ordinary particles is severely constrained.

Freezing Versus Thawing Model The quintessence field we have discussed so far is characterized by a field that is rolling toward its potential minimum prior to the onset of cosmic acceleration, but which slows down almost to a freezing level at present. There is another class of quintessence fields which shows the opposite character. The so-called thawing model has hardly moved in the past and has begun to roll down the potential recently [915, 916]. The initial slow motion is provided by the effective damping force of the Hubble expansion. If the field is slowly rolling at the beginning, the potential is very flat ($\dot{\phi}^2 \ll V$), which means the field has $w \simeq -1$ from the very beginning. A typical potential is of the form $V(\phi) = M^{4-n}\phi^n$, $n > 0$, which is ubiquitous (see Figure 11.10a). The power index n can be continuous. The mass of the field is given by $m_\phi^2 = V''$ at $V' = 0$, and both the mass and the potential are similar to those in the freezing model.

The slowness of the initial motion can be explained as follows: For a potential of the type $V = M^{4-n}\phi^n$ $(n > 0)$, the time derivatives $\dot{\phi}, \ddot{\phi}, V'$ can be approximated by $\phi/\Delta t, \phi/(\Delta t)^2, V/\phi$, where Δt is the duration of time over which the field can be considered as approximately constant. For a slow rolling field ($\ddot{\phi} \ll 3H\dot{\phi},\ V'$), the

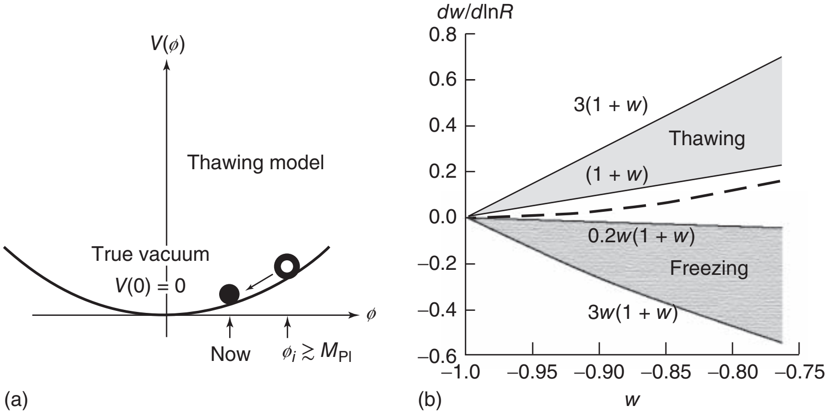


Figure 11.10 (a) Quintessence fields that hardly moved in the past but has begun to roll down the potential recently are called "the *thawing*" *models*. The approximate value of the potential energy is constrained by the present cosmic density ($V(\phi) \approx \rho(\phi) \simeq 10^{-30}$g cm$^{-3} \approx 10^{-120} M_{\text{Pl}}^4$). (b) The thawing model occupies an area different from that of the freezing model on the $w - w'$ plane and can be distinguished by observations [915].

field equation Eq. (11.15) means

$$\frac{3H\phi}{\Delta t} + \frac{V}{\phi} \approx 0 \quad \rightarrow \quad \Delta t \simeq \frac{1}{3H}\frac{V}{\left(\frac{\phi}{\Delta t}\right)^2} \sim \frac{1}{3H}\frac{V}{\dot{\phi}^2} \gg \frac{1}{3H}. \tag{11.42}$$

In other words, the field was frozen by the Hubble damping at a value displaced from its minimum until recently [denoted as ϕ_i in Figure 11.10a]. The initial field strength $|\phi_i|$ is at least of the order of or much larger than M_{Pl} because $\lambda = V'/(\kappa V) \sim M_{\text{Pl}}/\phi \ll 1$ [see Eq. (11.17)]. When the field begins to move and passes the point $\phi \sim M_{\text{Pl}}$, the first term in the field equation becomes comparable in size to the others and the field gains motion, evolving with time. As $dw/d\ln R \equiv \dot{w}/H > 0$, it is now moving to less negative values of w as the field rolls down toward the potential minimum at $\phi = 0$. The λ is no longer a small parameter, which means the cosmic acceleration stops. Therefore, the destiny of the Universe is very different for the thawing model. The acceleration is temporary as compared to the freezing field for which the acceleration continues unabated. Note that, unlike the freezing model which is insensitive to the initial condition but the potential form has to be chosen carefully, the thawing model needs a very carefully chosen initial condition while it is insensitive to the shape of the potential as long as the slow roll condition is satisfied.

As $w' \equiv dw/d\ln R = \dot{w}/H$ has the opposite sign and occupies different regions on the $w - w'$ plane (Figure 11.10b [915]) for the freezing and thawing models, they can be distinguished by observations.

Notice that the conditions for the quintessence field are very similar to those of the inflaton at the period of accelerating expansion in the early universe. Differences are that, unlike the inflaton, the dark energy need not be rolling as slowly and it is not totally the dominant component of the cosmic density.

11.4
Other Dark Energy Models

The quintessence model is the most orthodox model for the dynamic dark energy in the sense that it introduces only a new scalar field within the framework of the standard field theories. There are a variety of other scalar field models with differing potentials, such as the Doomsday model [917], the PNGB model [918], and so on. Other models introduce exotic ideas. The k-essence model [919, 920] has only kinetic terms and no potential terms. However, the kinetic energy has the form $A(\phi)X + B(\phi)X^2 + \cdots$, where $X = (1/2)g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi$ is the canonical kinetic energy term. The phantom model assumes a negative kinetic energy, which allows $w < -1$ [921]. The latter was motivated by the phenomenological observation that does not deny the possibility. The Chaplygin Gas model [922] assumes the equation of state $P = -A/\rho$ with A a positive constant. Some even modify the general relativity [923, 924] or rely on extra dimensions [925].

Cosmic Destiny In the old cosmology before the discovery of the dark energy, the geometry of the universe determined the ultimate destiny of the universe [891]. From the Friedmann equation Eq. (8.10) with $\Lambda = 0$ and the cosmic density-scale relation Eq. (8.53), we can express the cosmic expansion rate as

$$\dot{R}^2 = \Omega_0 H_0^2 R^{-(1+3w_T)} - k \tag{11.43}$$

where w_T is the equation of state parameter that takes values 0 (1/3) for matter (radiation) dominance. Therefore, in the absence of the dark energy, the first term on the right-hand side varies from ∞ to 0 as the universe expands. For $k > 0$, the expansion rate vanishes at some value of R where the scale factor takes a maximum value and shifts from expansion to contraction, eventually going to $R = 0$ (Big Crunch). For $k < 0$, Eq. (11.43) means eternal expansion. For $k = 0$, it means eternal expansion but its rate approaches zero. In short, an open, closed, or flat universe corresponding to $k <, =, > 0$ means uniquely eternal expansion, critical expansion, or recollapse.

The discovery of the dark energy, that is, the accelerating expansion, has changed the above picture drastically. Let us consider, first, the case in which the dark energy is the vacuum energy or, equivalently, the cosmological constant. In this case, $w = P_{\mathrm{DE}}/\rho_{\mathrm{DE}} = -1$ and the dark energy density remains constant, eventually dominating the cosmic energy. This is a case of eternal expansion but, unlike the old cosmology, the final destiny is very different. Consider a galaxy with receding speed v a little less or larger than the speed of light ($v = c \pm \varepsilon$). In the old cosmology in which the expansion is decelerating, the galaxy that had $v = c + \varepsilon$ now has $v < c$, and hence its light will reach us. It has entered the horizon.

On the other hand, if the expansion rate is accelerating, the one with $v = c - \varepsilon$ will soon be expanding at superluminal speed and will disappear beyond the (event) horizon. As time goes, more and more galaxies will disappear. Eventually, those beyond the local group ($r \geq 1 - 2$ Mpc) will be red-shifted beyond detectability on a timescale $t - t_0 \simeq 100$ Gyrs [926]. The Milky Way will remain gravitationally coupled to the local group which is left alone as an isolated island in the void of the vast universe [927, 928]. This is referred to as the *Red Out* [891].

What will happen if the dark energy is dynamic? If it is a scalar field, the field eventually settles at the minimum of its potential. If the minimum is precisely zero, the universe will become matter-dominated again and return to decelerating expansion, restoring the old cosmology scenario with the link between the geometry and destiny as stated before. If the minimum is negative, it means an eternal attraction leading to recollapse irrespective of the k value. If the potential stays positive, it is an eternal expansion to Red Out.

An interesting possibility exists for $w < -1$. It does not happen for the scalar field but it is conceivable if the field has negative kinetic energy dubbed as the "*phantom*" [921].

$$\text{Phantom Energy:} \quad \rho = -\frac{1}{2}\dot{\phi}^2 + V > 0, \quad P = -\frac{1}{2}\dot{\phi}^2 - V \tag{11.44}$$

From Eq. (11.25), we have

$$\rho = \rho_0 \left(\frac{R}{R_0}\right)^{\beta}, \qquad \beta = -3(1 + w) > 0. \tag{11.45}$$

The energy density increases with time. For simplicity, let us assume a flat universe and a constant equation of state parameter. Then, from the Friedmann equation and Eq. (11.45), we obtain

$$\begin{aligned} R &= \frac{R_0}{\left[1 - \frac{\beta H_0}{2}(t - t_0)\right]^{\frac{2}{\beta}}} = \left(\frac{2}{\beta H_0}\right)^{\frac{2}{\beta}} \frac{R_0}{(t_{\text{rip}} - t)^{\frac{2}{\beta}}} \\ t_{\text{rip}} &= t_0 + \frac{2}{\beta H_0}. \end{aligned} \tag{11.46}$$

The cosmic scale grows rapidly and reaches infinity at a finite time. The energy density also grows, eventually exceeding the binding energy of any structure. All structures, from clusters to atoms, will be torn apart. This is the scenario referred to as the *Big Rip* [929]. For $w = -5/3$, the Big Rip will happen after another 13.8 Gyrs.

11.5
How to Investigate the Dark Energy?

The Equation of State of the Dark Energy Although the data fit well with the ΛCDM model which assumes a constant cosmological constant, it could well be a result of some dynamical phenomenon or a new physics. In order to gain some insight into the nature of the dark energy, we discuss in this section how to extract its detailed information from observation, that is, the z dependence of the dark energy density $\Omega_{\mathrm{DE}}(z)$ or, equivalently, its equation of state parameter $w(z)$ [914, 930].

The time variation of the energy density is expressed as [Eq. (8.19)]

$$\dot{\rho} = -3H(\rho + P) = -3H(1+w)\rho. \tag{11.47}$$

The cosmological constant is a special set of ρs which can be obtained by setting the equation of state parameter $w = -1$. If w is a dynamical variable, the dark energy density can be cast in the form

$$\rho_{\mathrm{DE}}(z) = \rho_{\mathrm{DE}\,0}\exp\left[-\int_{t_0}^{t} dt\,3(1+w_{\mathrm{DE}})H(t)\right] = \rho_{\mathrm{DE}\,0}\exp\left[\int_{0}^{z} d\tilde{z}\frac{3(1+w_{\mathrm{DE}})}{1+\tilde{z}}\right]. \tag{11.48}$$

To compare the z dependence with that of matter or radiation, we introduce the average value $\hat{w}$ and rewrite Eq. (11.48) as

$$\rho_{\mathrm{DE}} = \rho_{\mathrm{DE}\,0}(1+z)^{3(1+\hat{w}_{\mathrm{DE}})}, \qquad 3(1+\tilde{w}_{\mathrm{DE}}) = \frac{1}{\ln(1+z)}\int_0^z d\tilde{z}\,\frac{w(\tilde{z})}{1+\tilde{z}}. \tag{11.49}$$

In order to derive the z dependence of the dark energy density, we need to express the Hubble parameter as a function of z.

$$\begin{aligned} H^2(t) = \left(\frac{\dot{R}}{R}\right)^2 &= \frac{8\pi G_N}{3}\rho - \frac{k}{R^2} = \frac{8\pi G_N}{3}\left(\rho_m + \rho_{\mathrm{rad}} + \rho_{\mathrm{DE}}\right) - \frac{k}{R^2} \\ &= H_0^2\left(\Omega_m + \Omega_{\mathrm{rad}} + \Omega_{\mathrm{DE}}\right) - \frac{k}{R^2} \end{aligned} \tag{11.50}$$

where $\Omega_i = \rho_i/\rho_{c0} = 8\pi G_N \rho_i/(3H_0^2)$ has been used. Taking into account the z dependence of each component, it can be expressed as

$$H^2(z) = H_0^2\left[\Omega_{\mathrm{DE}}(z) + \Omega_{\mathrm{rad}0}(1+z)^4 + \Omega_{m0}(1+z)^3 + \Omega_{K0}(1+z)^2\right] \tag{11.51a}$$

where

$$\Omega_{\mathrm{DE}}(z) = \frac{\rho_{\mathrm{DE}}}{\rho_{c0}} = \Omega_{\mathrm{DE}0}\exp\left[\int_0^z dz'\,\frac{3(1+w)}{1+z'}\right], \quad \Omega_{K0} = -\frac{k}{R_0^2H_0^2}. \tag{11.51b}$$

The suffix 0 denotes the values at the present time. It is convenient to define a function $E(z)$ by

$$E^2(z) \equiv \frac{H^2(z)}{H_0^2} = \Omega_{\mathrm{DE}}(z) + \Omega_{m0}(1+z)^3 + \Omega_{K0}(1+z)^2 \tag{11.52}$$

where we have set the radiation density equal to zero because the effect of the dark energy will be apparent only in the matter-dominated era. Differentiating Eq. (11.52) w.r.t. z, and using

$$\Omega'_{\mathrm{DE}}(z) \equiv \frac{d}{dz}\Omega_{\mathrm{DE}}(z) = \Omega_{\mathrm{DE}}(z)\frac{3(1+w_{\mathrm{DE}})}{1+z} \tag{11.53}$$

we obtain the z dependence of w_{DE} as

$$\begin{aligned}1+w_{\mathrm{DE}}(z) &= \frac{1+z}{3\Omega_{\mathrm{DE}}(z)}\left[\left\{E^2(z)\right\}' - 3\Omega_{m0}(1+z)^2 - 2\Omega_{K0}(1+z)\right] \\ &= \frac{1+z}{3}\frac{\left\{E^2(z)\right\}' - 3\Omega_{m0}(1+z)^2 - 2\Omega_{K0}(1+z)}{E^2(z) - \Omega_{m0}(1+z)^3 - \Omega_{K0}(1+z)^2}.\end{aligned} \tag{11.54}$$

The function $E(z)$ can be expressed in terms of an observable, namely the luminosity distance d_L [see Eq. (8.72)].

$$d_L = R_0 r(1+z) = R_0 S_k(\chi)(1+z) = \frac{1+z}{H_0\sqrt{\Omega_{K0}}}\sinh\left[\sqrt{\Omega_{K0}}\int_0^z \frac{d\tilde{z}}{E(\tilde{z})}\right] \tag{11.55}$$

Using Eq. (11.55), Eq. (11.52) can be rewritten as

$$E^2(z) = \frac{(1+z)^2\left[(1+z)^2 + \Omega_{K0}H_0^2 d_L^2(z)\right]}{\left[(1+z)H_0 d_L{}'(z) - H_0 d_L(z)\right]^2}. \tag{11.56}$$

Equation (11.56) tells that one can determine the function $E(z)$ by measuring the z dependence of the luminosity distance and, by extension, the angular diameter distance, too. The angular diameter distance is simply related to the luminosity distance by $d_A = d_L/(1+z)^2$. Figure 11.11 shows the z dependence of the distances in units of the inverse Hubble distance cH_0^{-1}.

For a flat universe ($\Omega_K = 0$), for which $d_L = \frac{1+z}{H_0}\int_0^z \frac{d\tilde{z}}{E(\tilde{z})}$, Eq. (11.56) is simplified to

$$E(z) = \frac{1}{H_0}\left[\frac{d}{dz}\left(\frac{d_L(z)}{1+z}\right)\right]^{-1} \qquad : \text{flat universe} \tag{11.57}$$

and Eq. (11.54) reduces to

$$1+w_{\mathrm{DE}}(z) = \frac{1+z}{3}\frac{\left[\left\{E^2(z)\right\}' - 3\Omega_{m0}(1+z)^2\right]}{\left[E^2(z) - \Omega_{m0}(1+z)^3\right]} \qquad : \text{flat universe.} \tag{11.58}$$

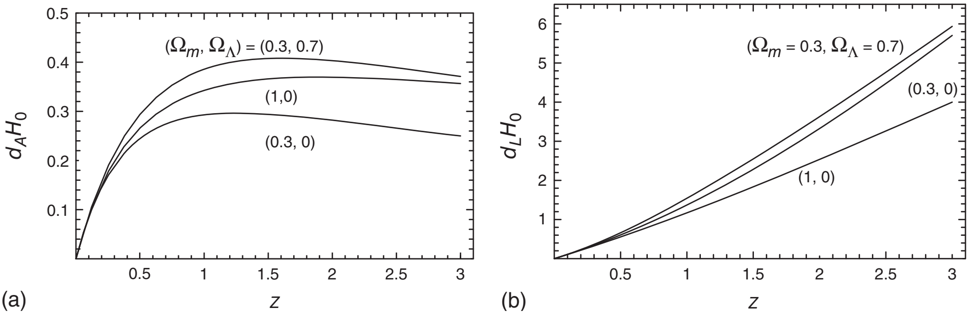


Figure 11.11 The angular diameter distance in units of cH_0^{-1} as a function of the red shift is shown in (a). The luminosity distance in units of cH_0^{-1} as a function of the red shift is shown in (b). Each curve is labeled by $(\Omega_m, \Omega_\Lambda)$ (Reproduced with permission of [931].)

One can also determine directly the dark energy density $\rho_{\mathrm{DE}}(z)$ by

$$\rho_{\mathrm{DE}}(z) = \rho_{c0}\Omega_{\mathrm{DE}}(z) = \frac{3H_0^2}{8\pi G_N}\left[E^2(z) - \Omega_{m0}(1+z)^3\right] \qquad : \text{flat universe.} \quad (11.59)$$

If the dark energy is due to a single rolling scalar field, its potential can also be reconstructed.

$$\begin{aligned} V\bigl(\phi(z)\bigr) &= \frac{1}{2}\rho_{\mathrm{DE}}(z)[1 - w_{\mathrm{DE}}(z)] \\ &= \frac{3H_0^2}{16\pi G_N}\left[2E^2(z) - \frac{1+z}{3}\{E^2(z)\}' - \Omega_{m0}(1+z)^3\right] \qquad : \text{flat universe} \end{aligned} \quad (11.60)$$

where Eqs. (11.58) and (11.59) were used to derive the formula.

In practical analyses, however, it is difficult to use the above exact formula to directly reconstruct $w_{\mathrm{DE}}(z)$ *or* $\rho_{\mathrm{DE}}(z)$, and one often uses a linearized approximation. In fitting the data with the equation of state parameter, $w(z)$ is conventionally parameterized as

$$w = w_0 + w_1 z, \qquad \text{or} \quad w(R) = w_0 + w_a\left(1 - \frac{R}{R_0}\right) = w_0 + w_a\frac{z}{1+z}. \quad (11.61)$$

The second form is useful for treating data with large z.

Observables We have argued that, in principle, the complete z dependence of the dark energy ρ_{DE} and w_{DE} can be determined by measuring the z dependence of the luminosity distance d_L and/or the angular diameter distance d_A. Figure 11.11 shows the z dependence of the distances in units of cH_0^{-1}. The effect of the w variable causes a slight change in the relation between the angular or luminosity distance and the red shift and modifies the spectrum of observables.

Another important geometrical variable that is sensitive to w is the cosmic volume element which is necessary in observing the z dependence of the number counts, for example, counts of lensed quasars, galaxies, or cluster of galaxies. Denoting $n(z)$ as the number density per unit comoving volume, one can express the observed number of clusters of galaxies in a survey of solid angle Ω_{survey} centered at the red shift z and in a shell of thickness Δz as

$$\begin{aligned} N(z,\Delta z) &= \Omega_{\mathrm{survey}}\int_{\frac{z-\Delta z}{2}}^{\frac{z+\Delta z}{2}} n(z)\frac{dV}{dzd\Omega}dz \\ \frac{dV}{dzd\Omega} &= \frac{1}{dzd\Omega}\frac{r^2 drd\Omega}{\sqrt{1-kr^2}} = \frac{r^2}{H(z)} = \frac{1}{H(z)}\left[(1+z)d_A\right]^2. \end{aligned} \quad (11.62)$$

The ability of these cosmological observables to probe the dark energy equation of state depends upon their sensitivity to w. Figure 11.12a shows the relative changes in $dV/dzd\Omega$, $r(z)$ and $H(z)$ [932]. The sensitivities peak around $z \sim 1$ and level off at higher z. Observations at red shifts $0 \lesssim z \lesssim 2$ will be most useful in probing the dark energy.

Figure 11.12b shows the predicted cluster counts for a survey covering 4000 sq. deg. which is sensitive to halos more massive than $2 \times 10^{14}\, M_\odot$ for three flat

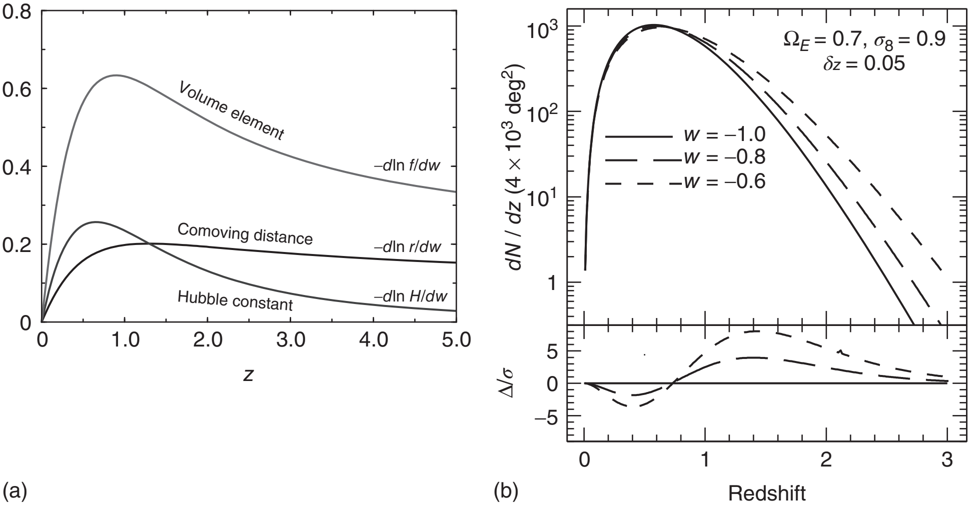


Figure 11.12 (a) Relative sensitivity of $f(z) = dV/dzd\Omega$, $r(z)$, and $H(z)$ to a change in the dark energy equation of state. [932] (b) Predicted cluster counts for a survey covering 4000 sq. deg. which is sensitive to halos more massive than $2 \times 10^{14}\ M_\odot$, for three flat cosmological models with fixed $\Omega_m = 0.3$ and $\sigma_8 = 0.9$. The lower panel shows the fractional differences between the models in terms of the estimated Poisson errors (Reproduced with permission of [891, 933, 934].)

cosmological models. The lower panel shows the fractional differences between the models [891, 933, 934].

Among the observables that use the luminosity distance are SNIa and GRB (gamma ray burst), and those that use the angular diameter distance are CMB and gravitational lensing. The large scale structure (LSS) is sensitive to volume elements.

As an example of the use of the angular diameter distance, we show in Figure 11.13a a CMB spectrum for three values of w [935].[9]

As an example for the use of the luminosity distance, we show in Figure 11.13b the predicted distance modulus [see Eq. (9.8) for its definition] of type SNIa as a function of z as predicted by SNAP (the supernova acceleration probe).[10]

Notice that the distance modulus is the apparent brightness magnitude, which is proportional to the logarithm of the luminosity distance. The ordinate of Figure 11.13b is normalized to the ΛCDM model with $w = -1$. The graph shows that it can discriminate between different cosmological models with different constants w as well as quintessence models (two representative models (SUGRA) and (two EXP) are shown).

Figure 11.14 summarizes the current constraints from collections of SNe, CMB, and BAO data [906].

---

9) There is a degeneracy problem. The same curve can be reproduced by using a different set of parameter values. To resolve the degeneracy, one has to combine CMB data with other sets of observables.

10) It was a part of JDEM (joint dark energy mission) [939] and now a part of WFIRST [940].

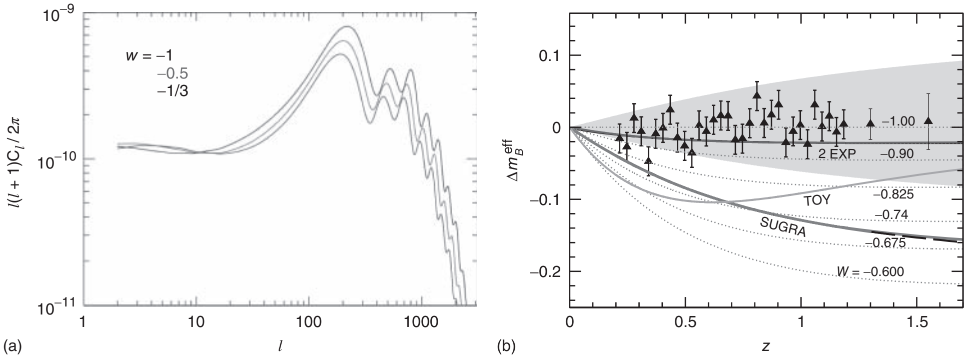


Figure 11.13 (a) Example of w dependence of the CMB spectra. (Reproduced with permission of [935].) (b) Simulated data using SNAP (supernova acceleration probe), with solid curves from a supergravity-inspired quintessence model [936] (with the best $N = 1$ fit shown as a dashed line, mostly covered by the solid SUGRA line), a strongly evolving toy model with $w_0 = -0.6$ and $w_1 = -0.8$ (TOY), and a quintessence model involving two exponentials (2EXP) [937]. A range of constant w models is shown as a grid of lines. The shaded region corresponds to cosmological constant models with $0.25 < \Omega_m < 0.35$ and $\Omega_\Lambda = 1 - \Omega_m$ (Reproduced with permission of [938].)

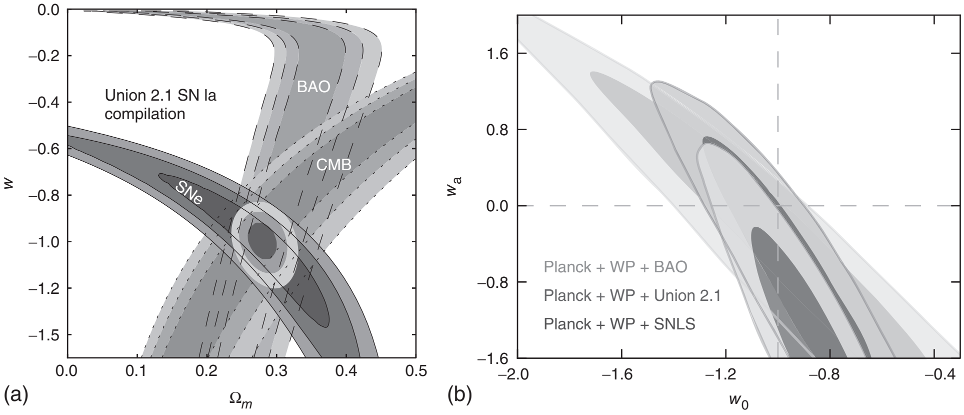


Figure 11.14 (a) 68.3%, 95.4%, and 99.7% confidence regions of the (Ω_m, w) plane. (Reproduced with permission of [905, 906].) (b) 2-D marginalized posterior distributions for w_0 and w_a, for the data combinations Planck+WP+BAO (gray), Planck+WP+Union2.1 (red) and Planck+WP+SNLS (blue). The contours are 68% and 95%, and dashed gray lines show the cosmological constant solution (Reproduced with permission of [123].) (Please find a color version of this figure on the color plates.)

The present value is consistent with w being a constant, whose value is [123]

$$w = -1.13^{+0.23}_{-0.25} \qquad \begin{cases} w_0 & = -1.04^{+0.72}_{-0.69} \\ w_a & < \quad 1.32. \end{cases} \tag{11.63}$$

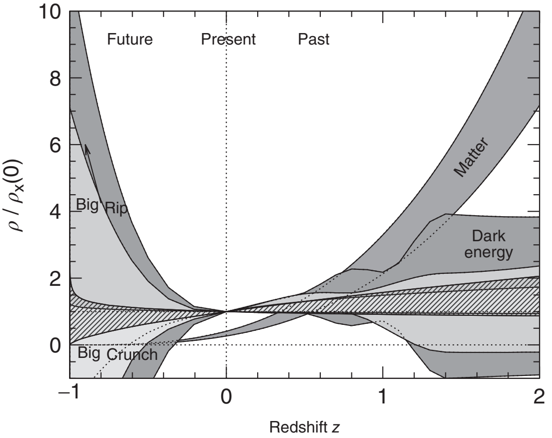


Figure 11.15 1σ constraints on the density of matter and dark energy from SN Ia, CMB, and LSS data, all in units of the current dark energy density (Reproduced with permission of [941].) From inside out, the four nested dark energy constraints are for models making increasingly strong assumptions, corresponding, respectively, to the four-parameter spline, the three-parameter spline, the two-parameter f_∞, w_i case, and the one-parameter constant w case (hatched area). The Universe starts accelerating when the total density slope $d\ln\rho/d\ln(1+z) > -2$, which roughly corresponds to when the dark energy begins to dominate, that is, to where the matter and dark energy bands cross. In the distant future, the Universe will recollapse if the dark energy density ρ_X goes negative and ends in a "Big Rip" if it keeps growing ($d\ln\rho_X = d\ln(1+z) < 0$). (Please find a color version of this figure on the color plates.)

Figure 11.15 shows an attempt to extract the z dependence of the dark energy density [941].

Future Projects Currently, investigations of the dark energy are being carried out by three complementary methods, observation of supernovae Ia (SNIa), baryon acoustic oscillation (BAO), and gravitational lensing (GL or WL for weak lensing) complemented by CL (cluster distributions) and RSD (Red Shift-Space Distortions which measures effects of cosmic acceleration on structure growth). SN and BAO measure the distances as a function of the red shift z. CL measures the abundance of the most massive galaxy clusters, while RSD measures the flows of matter. GL is hybrid because lensing depends on both the distance to the lensing mass (distance test) and the amount of lensing mass (structure growth test).

As the w dependences of the distances (d_L, d_A) are more conspicuous at $z > 1$, we want observables at large red shifts like Lyα[11] and GRBs. The Lyα information is part of the large-scale structure data. We also want to use GRB extensively. Unfortunately, it has no established status as a standard candle because there is

---

[11] The Lyman alpha spectrum of hydrogen atoms emitted by a quasar at large z_Q shows many dips (absorption lines) at various $z < z_Q$ (referred to as the *Lyman α forest*), which represent neutral hydrogen cloud (intergalactic gas) distribution in the light path.

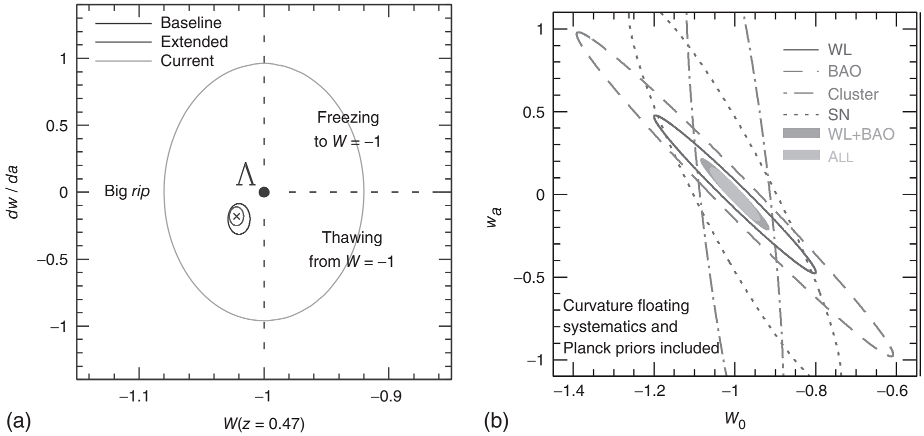


Figure 11.16 (a) Forecast constraints on dark energy parameters from WFIRST-2.4 compared to current knowledge. *dw/da* is $-w_a$ in the text. Ellipses show $\Delta\chi^2 = 1$ error contours on the value of w at redshift z = 0.47. The green ellipse, centered here on the cosmological constant model (w = -1, dw/da = 0), represents current state-of-the-art constraints from a combination of CMB, SN, BAO, and H0 data, based on Anderson *et al.* (2012). The black ellipse shows the error forecast for the baseline WFIRST-2.4 supernova, galaxy redshift, and weak lensing surveys, combined with CMB data from Planck, a local supernova calibrator sample, and measurements of BAO and redshiftspace distortions from the SDSS-III BOSS survey at $z < 0.7$. The red ellipse shows the "extended" case in which the precision of the WFIRST-2.4 is increased by a factor of two, as a result of a longer observing program. (Reproduced with permission of [940, 944].) (b) Forecasts of constraints on $w_0 - w_a$ region. LSST BAO (baryon acoustic observation, dashed line), cluster counting (dash-dotted line), supernovae (dotted line), WL (weak lensing, solid line), joint BAO and WL (green shaded area), and all combined (yellow shaded area). The BAO and WL results are based on galaxy–galaxy, galaxy–shear, and shear–shear power spectra only. Adding other probes such as strong lensing time delay and higher order galaxy and shear statistics will further improve the constraints (Reproduced with permission of [945, 946].) (Please find a color version of this figure on the color plates.)

no other independent means to cross-check its validity, and its dispersion is wide compared to SNe.

Several new projects are already in operation and are providing more accurate data. On CMB, Atacama Cosmology Telescope (ACT) [942] and the South Pole Telescope (SPT) are ground-based projects [124,943]. The Planck Satellite, launched in May 2009, is publishing data on high-precision all-sky maps of temperature and polarization [123]. It will provide more accurate data than the WMAP.

There are many future projects to investigate the dark energy further. We only list a few, which are the most advanced (or long-term ~2020 – 2022).[12] For space-based projects, we mention EUCLID, a project of ESA (the European Space Agency) and AFTA/WFIRST (Astrophysics Focused Telescope Assets/Wide Field Infrared Survey Telescope) proposed to NASA and DOE [944]. The latter uses an NIR (near infrared) 2.4-m telescope as its main instrument with image field coverage 200 times larger than the Hubble Space Telescope. It will do so at the resolution of HST. It can sample millions of galxies in the redshift range $1 < z < 3$ with $\Delta z/z \sim 1\%$.

---

[12] Be cautious however, many projects have or being amalgameted and change their project names before they can be realized. This is due to rapid progress in the field, developement of new detector devices and cost savings.

LSST (Large Synoptic Survey Telescope) is a large, wide-field, ground-based optical/infrared telescope (8.4 m in diameter) designed to obtain multiband images over a substantial fraction of the sky [945]. The LSST project uses the combined techniques of WL/CL/SN/BAO to obtain comprehensive data. It is to be sited at Cerro Pachon in Chile. MS-DESI (Mid-Scale Dark Energy Spectroscopic Instrument) is the unification of the two formerly competing BigBOSS and DES and is considered as an intermediate project connecting the present SDSS/Boss to LSST.

Figure 11.16 shows expected achievements of WFIRST and LSST which are hoped to be realized in about ten years.

The quest for the dark energy is still in its infancy, and accumulation of data is an absolute must in order for us to have more insight into it.

Appendix A
Virial Theorem

The virial theorem is a general equation in mechanics between the average total kinetic energy $<K>$ of a stable system consisting of N particles bound by a conservative force and the potential $<U>$ where the brackets mean the average over time. Denoting the mass, position, velocity, and momentum vectors of the kth particle as m_k, $\mathbf{r}_k$, $\mathbf{v}_k = \dot{\mathbf{r}}_k$, and $\mathbf{p}_k$, we start from the moment I defined by

$$I = \frac{1}{2}\sum_{k=1}^{N} m_k |\mathbf{r}_k|^2, \tag{A.1a}$$

$$\frac{dI}{dt} = \sum m_k \mathbf{r}_k \cdot \dot{\mathbf{r}}_k = \sum \mathbf{r}_k \cdot \mathbf{p}_k, \tag{A.1b}$$

$$\frac{d^2 I}{dt^2} = \sum \left[\dot{\mathbf{r}}_k \cdot \mathbf{p}_k + \mathbf{r}_k \cdot \frac{d\mathbf{p}_k}{dt}\right] = 2K + \sum \mathbf{r}_k \cdot \mathbf{F}_k = 2K - \sum \mathbf{r}_k \cdot \nabla_k U. \tag{A.1c}$$

If the potential is a homogeneous function of degree n, the following equality should hold:

$$U(a\mathbf{r}_1, a\mathbf{r}_2, \cdots) = a^n U \quad \Rightarrow \quad \sum \mathbf{r}_k \cdot \nabla_k U(\mathbf{r}_1, \mathbf{r}_2, \cdots) = nU. \tag{A.2}$$

Substituting Eq. (A.2) into (A.1c)

$$\frac{d^2 I}{dt^2} = 2K - nU. \tag{A.3}$$

Now we take the average over a long time τ. Then the LHS becomes

$$\left\langle \frac{d^2 I}{dt^2} \right\rangle = \frac{1}{t}\left\langle \left.\frac{dI}{dt}\right|_t - \left.\frac{dI}{dt}\right|_0 \right\rangle, \qquad \text{Note : F } < (\cdots) >= \frac{1}{\tau}\int_0^\tau dt(\cdots). \tag{A.4}$$

If the system is cyclical such that it returns to its initial state after an interval τ, the above quantity vanishes. If it is not cyclic, for a system being bounded, the limit vanishes, too. For a random system, the first and second terms can be considered to cancel each other. Another way of thinking is that the moment of the multi-particle system takes a steady value after a long time, which is referred to as the *virialized state*. Therefore, we have the equality

$$< K >= \frac{1}{2} n < U > . \tag{A.5}$$

Combining with the energy conservation law

$$E = K + U \tag{A.6}$$

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

we have

$$< K >= \frac{n}{n+2} <E>, \qquad < U >= \frac{2}{n+2} <E> \tag{A.7}$$

After a long time, and if N is sufficiently large, the system can be considered to be in thermal equilibrium. Denoting the temperature as T

$$\frac{3}{2}kT = \sum_{k=1}^{3} \frac{1}{2} < m_k v_k^2 >= \frac{< K >}{N} = \frac{n < U >}{2N} = \frac{n}{2} < u > . \tag{A.8}$$

where $<u>= U/N$ indicates the average potential energy of each particle. The temperature defined by (A.8) is referred to as the *virial temperature*. It can be applied to a system that is not completely in thermal equilibrium. For the gravitational potential, $n = -1$, so

$$< K >= -\frac{1}{2} < U >= -E. \tag{A.9}$$

Appendix B
Chandrasekhar Mass

Polytrope gas: When the interior of a star is in dynamical equilibrium, the mass and the momentum conservation laws can be written as

$$\frac{dM(r)}{dr} = 4\pi r^2\rho, \qquad \frac{dP}{dr} = -G_N\frac{M(r)\rho}{r^2} \tag{B.1}$$

which have three independent variables $M(r)$, $P(r)$, and $\rho(r)$. Therefore, if one assumes a relation

$$P(r) = K\rho^{\gamma} \equiv K\rho^{1+\frac{1}{N}} \tag{B.2}$$

where γ is the adiabatic index and K is a constant, one can solve the equations. A sphere of gas satisfying Eq. (B.2) is referred to as a *polytrope gas* and is used as a model in many circumstances discussing the properties and dynamical conditions of the main sequence stars. Especially, it is known that the properties of the Sun is well reproduced by an $N = 3$ polytrope. As a completely degenerate gas sphere (which is our interest) has $\gamma = 5/3$ (nonrelativistic) or $\gamma = 4/3$ (extreme relativistic), it is well approximated by a polytrope gas.

Notice that Eq. (B.2) merely gives the condition between the pressure and the density. It is not an equation of state. The latter includes an additional parameter, which is temperature T. But if temperature comes in, the number of independent variables increases to 4 and one needs an additional equation which is supplied by the energy transport equation. However, if the star's interior is in convective equilibrium, then the energy transport process is adiabatic and $N = 1/(\gamma - 1)$ holds.

From Eq. (B.1), one obtains

$$-G_N\frac{dM(r)}{dr} = \frac{d}{dr}\left(\frac{r^2}{\rho}\frac{dP}{dr}\right) = -4\pi G_N r^2\rho(r). \tag{B.3}$$

Introducing new parameters θ_N and ξ by

$$\rho(r) = \rho_c\theta_N(\xi)^N, \quad r = \alpha\xi \tag{B.4a}$$

$$\alpha = \left[\frac{(N+1)}{4\pi G_N}\frac{P_c}{\rho_c^2}\right]^{\frac{1}{2}} \tag{B.4b}$$

where P_c and ρ_c denote the pressure and density at the core, we obtain the equation

$$\frac{1}{\xi^2}\frac{d}{d\xi}\left(\xi^2\frac{d\theta_N}{d\xi}\right) = -(\theta_N)^N. \tag{B.5}$$

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

Equation (B.5) is referred to as the *Lane--Emden* equation of index N. As the density is finite and the pressure gradient should vanish at the center, we have boundary conditions

$$\theta_N(0) = 1, \qquad \theta'(0) = 0. \tag{B.6}$$

To solve Eq. (B.5) for general N, one has to rely on numerical calculations. But for the special values of $N = 0, 1, 5$, analytical solutions are known.

$$\begin{aligned} \theta_0(\xi) &= 1 - \frac{1}{6}\xi^2, \\ \theta_1(\xi) &= \frac{\sin\xi}{\xi}, \\ \theta_5(\xi) &= \frac{1}{[1 + (\xi^2/3)]^{\frac{1}{2}}}. \end{aligned} \tag{B.7}$$

The $N = 5$ solution does not vanish for a finite value of ξ; hence its radius is infinite. Since the density must be positive ($\theta_N(\xi) > 0$), Eq. (B.5) constrains

$$\frac{d\theta_N}{d\xi} = -\frac{1}{\xi^2}\int \xi^2(\theta_N)^N(\xi)d\xi < 0. \tag{B.8}$$

Therefore, $\theta_N(\xi)$ is a monotonically decreasing function of ξ. The first zero point corresponds to surface of the star, which is denoted as ξ_N. For $N < 5$, one can derive from Eq. (B.4)

$$R_N = \alpha\xi_N = \left[\frac{N+1}{4\pi G_N}\frac{P_c}{\rho_c^2}\right]^{\frac{1}{2}}\xi_N \tag{B.9a}$$

and the mass of the star

$$M = \int 4\pi\rho r^2 dr = -4\pi\alpha^3\rho_c\left[\xi^2\frac{d\theta_N}{d\xi}\right]_{\xi=\xi_N} = \left[\frac{1}{4\pi G_N^3}\frac{P_c^3}{\rho_c^4}\right]^{\frac{1}{2}}\varphi_N \tag{B.9b}$$

where

$$\varphi_N = -(N+1)^{\frac{3}{2}}\left[\xi^2\frac{d\theta_N(\xi)}{d\xi}\right]_{\xi=\xi_N}. \tag{B.9c}$$

The values of φ_N is given in Table B.1. As the dependence of φ_N on N is not large, we may conveniently regard it almost constant.

Solving Eqs. (B.9) inversely, one obtains

$$\rho_c = \frac{M}{4\pi R_N^3}\frac{\beta_N^3}{\varphi_N} = \bar{\rho}\frac{\beta_N^3}{3\varphi_N}, \qquad \beta_N = (N+1)^{\frac{1}{2}}\xi_N, \tag{B.10a}$$

$$P_c = \frac{G_N M^2}{4\pi R_N^4}\frac{\beta_N^4}{\varphi_N^2} = (4\pi)^{\frac{1}{3}}G_N\left(\frac{M\rho_c^2}{\varphi_N}\right)^{\frac{2}{3}}. \tag{B.10b}$$

The second equality for P_c can be obtained from ρ_c by removing R_N. These equalities mean that the density and pressure at the core can be deduced from the stellar mass and its radius. In particular, the dependence on N of P_c is in φ_N, which is a slowly varying function of N and the expression for P_c is almost universal. It is interesting

Table B.1 Examples of polytrope gas.

γ	N	$(N+1)^{1/2}\xi_N$	φ_N	$\rho_c/\bar{\rho}$
1	∞[a]	∞	∞	∞
6/5	5	∞	25.46	∞
5/4	4	33.48	20.09	6.19×10^3
4/3	3[b]	13.79	16.15	54.2
7/5	2.5	10.02	14.32	23.4
3/2	2	7.539	12.53	11.4
5/3	1.5[c]	5.777	10.73	5.99
2	1	4.443	8.886	3.29
∞	0[d]	$\sqrt{6}$	$2\sqrt{6}$	1

[a] Equithermal gas sphere.
[b] Extreme relativistic gas sphere (white dwarf, super massive star).
[c] Nonrelativistic gas sphere (white dwarf, ordinary stars).
[d] Noncompressible gas sphere.

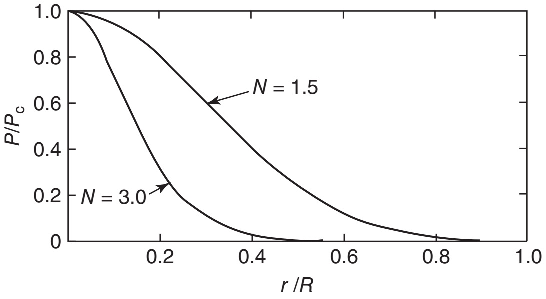


Figure B.1 Pressure distributions of a polytrope gas with index $N = 1.5$ and 3.

to note that $P_c \propto \rho_c^{4/3}$ regardless of N (or adiabatic index), since the adiabatic index γ changes by nuclear reactions and chemical compositions. Figure B.1 shows the pressure distributions for a polytrope gas with $N = 1.5$ and 3. The two models are simple but represent the characteristics of the stellar interior fairly well.

By removing P_c and ρ_c from the first equalities of Eq. (B.10) and Eq. (B.2) , one obtains

$$R_N^{3-N} M^{N-1} = \frac{K^N \varphi_N^{N-1} \beta_N^{3-N}}{4\pi (G_N)^N} = K^N D^N \tag{B.11}$$

D_Ns are known constants. That is, for $1 < N < 3$ (or $4/3 < \gamma < 2$)

$$R^{3-N} \propto \frac{1}{M^{N-1}}. \tag{B.12}$$

This means that the stellar radius becomes smaller as its mass becomes larger. When N is given and if K cannot be varied from the constraints of the equation of state, the constraint Eq. (B.11) forces the mass M or R to be fixed when one of them is specified. When K is not a free parameter, the number of independent inputs is limited to 1.

Assuming that the Sun is a polytrope gas sphere with index $N = 3$, and knowing $M_\odot = 2 \times 10^{30}$ kg and $R_\odot = 7 \times 10^8$ m, one obtains

$$\begin{aligned} \rho_c &= 75.3\,\mathrm{g\,cm^{-3}} \\ P_c &= 1.23 \times 10^{16}\ \mathrm{N\,m^{-2}} \\ T_c &= 1.2 \times 10^7\ \mathrm{K} \end{aligned} \tag{B.13}$$

where $\mu = 0.62$ (the average number of nucleons per electron) was assumed to derive the core temperature. The above numbers are accurate within 30 %.

Chandrasekhar mass: Stars like white dwarfs or neutron stars are supported by degenerate gas pressure against the gravitational collapse, the former by electrons and the latter by neutrons. For completely degenerate relativistic electrons, the number density and the pressure are given by

$$n_e = \int dn = 2\int_0^{p_F} \frac{4\pi p^2 dp}{(2\pi)^3} = \frac{p_F^3}{3\pi^2}, \tag{B.14a}$$

$$P_c = \frac{1}{3}\int vpdn = \frac{p_F^4}{12\pi^2} = \frac{(3\pi^2 n_e)^{\frac{4}{3}}}{12\pi^2} = \frac{1}{12\pi^2}\left(\frac{3\pi^2}{m_N \mu_e}\rho_c\right)^{\frac{4}{3}} \equiv K\rho_c^{\frac{4}{3}} \tag{B.14b}$$

where μ_e is the average number of nucleons per electron. One sees that Eq. (B.14b) satisfies the polytrope condition with $\gamma = 4/3$, *i.e.* N=3. Substituting Eq. (B.14b) and the value of φ_3 in Eq. (B.9b), we can calculate mass of the white dwarfs or degenerate core of evolving stars.

$$M = \left[\frac{1}{4\pi G_N^3}\frac{P_c^3}{\rho_c^4}\right]^{\frac{1}{2}} \varphi_3 = \frac{\sqrt{3\pi}}{16}\left(\frac{1}{G_N^3 m_N^4 \mu_e^4}\right)^{\frac{1}{2}} \varphi_3 = 1.46\left(\frac{2}{\mu_e}\right)^2 M_\odot \equiv M_{\mathrm{Ch}}. \tag{B.15}$$

This is referred to as the *Chandrasekhar mass*.

Appendix C
Production of KK Gravitons

***F_i* and *G_i* functions:** The $F_i(x, y)$ functions are [467]

$$F_1(x,y) = \frac{1}{x(y-1-x)}\left[-4x(1+x)(1+2x+2x^2) + y(1+6x+18x^2+16x^3) - 6y^2x(1+2x) + y^3(1+4x)\right], \tag{C.1a}$$

$$F_2(x,y) = -(y-1-x)F_1\left(\frac{x}{y-1-x}, \frac{y}{y-1-x}\right), \tag{C.1b}$$

$$F_3(x,y) = \frac{1}{x(y-1-x)}\left[1+2x+3x^2+2x^3+x^4 - 2y(1+x^3) + 3y^2(1+x^2) - 2y^3(1+x) + y^4\right]. \tag{C.1c}$$

The function $F_1(x, y)$ determines the cross section for $f\bar{f} \to \gamma G$. Because of charge conjugation invariance in Quantum ElectroDynamics, $F_1(x, y)$ is invariant under exchange of the Mandelstam variables t and u. This is reflected in the property $F_1(x, y) = F_1(y - 1 - x, y)$. The same property holds for the function F_3, which is relevant to the QCD process. The G_i functions are

$$G_1(x) = \left[\frac{1+2x+2x^2}{-x(1+x)}\right]^{\frac{1}{2}}, \tag{C.2a}$$

$$G_2(x) = \left[-\frac{x}{16}(1+x)(1+2x+2x^2)\right]^{\frac{1}{2}}, \tag{C.2b}$$

$$G_3(x) = 1+4x+6x^2+4x^3+2x^4, \tag{C.2c}$$

$$G_4(x) = 1+10x+42x^2+64x^3+32x^4, \tag{C.2d}$$

$$G_5(x) = 1+6x+12x^2+8x^3, \tag{C.2e}$$

$$G_6(x) = 1+6x+6x^2, \tag{C.2f}$$

$$G_7(x) = 9x^{-1}+22+24x+11x^2+x^3, \tag{C.2g}$$

$$G_8(x) = 4+9x+6x^2+x^3, \tag{C.2h}$$

$$G_9(x) = 9+18x+15x^2+5x^3, \tag{C.2i}$$

$$G_{10}(x) = 1+12x+15x^2+5x^3, \tag{C.2j}$$

$$G_{11}(x) = 40+114x+126x^2+60x^3+9x^4. \tag{C.2k}$$

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

The functions $G_1 \sim G_6$ are also invariant under exchange of t and u and therefore satisfy the equality $G_i(-1-x) = G_i(x)$.

Appendix D
Homotopy

Homotopy: Solutions $\phi^a(x_\mu)$ to a field equation of motion where the field possesses n degrees of internal freedom $[\phi^a\ (a = 1, 2, \cdots, n)]$, can be considered as a projection from a four-dimensional spacetime to n-dimensional ϕ space. If the field ϕ is a continuous function of x_μ and when x_μ moves continuously from a point P to another point Q, the field ϕ also moves from a point A in ϕ space to another point B. If the path P → Q changes, the corresponding path A → B changes also (see Figure D.1a). Generally, when two different paths with the same end points can be made to overlap completely by continuous deformation of the paths, we say that they belong to the same homotopy class or that they are homotopic. The function that operates on the path and deforms (transforms) the path continuously is referred to as the *homotopy*. When the ϕ space has a special structure, for instance, one like Figure D.1b, there are paths that cannot be brought to overlap completely by continuous transformations. ϕ and ψ on the Figure D.1b are such examples. Then we say they belong to different homotopy classes. If they represent vacuum state of the field, they cannot be differentiated physically. If more than one degenerate vacua exist belonging to different homotopy classes, they provide conditions for the solitons to exist.

We need to solve the equation of motion to prove the existence of the soliton solution. However, it helps to learn mathematics of homotopy to understand necessary conditions for existence of the solitons in general and know how to classify them.

S^1 and R^1: Representation matrices of a Lie group G can be parameterized by variables ω^a. The space that the variables ω's span is referred to as a manifold. When its volume is finite, the group is called a *compact group*.

Example 1: Collection of real numbers R^1 makes a group with operation of addition and subtraction. Identity operation is addition of zero and inverse of addition (subtraction) is subtraction (addition) of the same number. The range of the variables $(-\infty < x < \infty)$ is infinite and R^1 is not compact.

Example 2: A phase transformation $U(1)$ of one-dimensional complex variable space or a rotation $SO(2)$ in two-dimensional real number space can be parameterized by rotation angle θ. It has a limited range $0 \le \theta < 2\pi$ hence they are

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

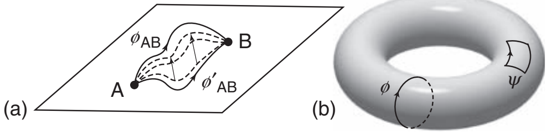


Figure D.1 Classification of projections. There are several paths to connect ϕ_A to ϕ_B depending on how x_μ moves in space from P to Q. (a) Two paths ϕ_{AB} and ϕ'_{AB} in the ϕ space can be continuously deformed (transformed) to make them overlap completely. (b) Paths of ψ and ϕ cannot overlap to each other by continuous deformation of their paths.

compact. The volume of their manifold is 2π. It has the same topology as a circle (S^1 a sphere in two-dimensional space). Now consider a real number x on a line ($A \leq x < B$). If we identify the points A and B, it is possible to transform x to θ by a continuous function $f(x)$ hence they are said to have the same topology. Setting A and B equal to $\pm\infty$, it becomes R^1 with points at infinity identified. Therefore, the R^1 with this condition is equivalent to S^1. Similarly, a two-dimensional plane R^2 (or three-dimensional cube R^3) when their points at infinity are identified, become topologically equivalent to $S^2(S^3)$ in two-(three-)dimensional spheres (or surface of sphere embedded in three-(four-)dimensional real space).

Connectedness: Let us consider relations between $SU(2)$ and $SO(3)$ both of which are referred to as the *rotation*. The transformation matrix of $SU(2)$ is expressed as

$$\begin{aligned} U &= \exp\left(-i\alpha\frac{\boldsymbol{\tau}\cdot\mathbf{n}}{2}\right) = \cos\frac{\alpha}{2} - i\boldsymbol{\tau}\cdot\mathbf{n}\sin\frac{\alpha}{2} \\ &\equiv a + \mathbf{b}\cdot\boldsymbol{\tau}, \qquad a^2 + b_1^2 + b_2^2 + b_3^2 = 1 \end{aligned} \tag{D.1}$$

The equation represents (S^3), a surface of a supersphere embedded in four-dimensional real space. Obviously any closed path on the surface of the sphere can be deformed continuously to a point. We say that the group $SU(2)$ is simply connected and is denoted mathematically as

$$\pi_1(SU(2)) = 0 \text{ [1]} \tag{D.2}$$

$\pi_1(SU(2))$ denotes the homotopy classification of the projection from S^1 to $SU(2)$ manifold and that all the $SU(2)$ projections belong to one homotopy class. Now let us consider $SO(3)$ which represents a rotation in three-dimensional real space. Using the Euler angles, it can be parameterized as

$$\begin{aligned} R &= e^{+\alpha J_z} e^{+i\theta J_y} e^{i\beta J_z} \\ &= \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{D.3}$$

---

[1] In fact, any sphere S^n with $n \geq 2$ is simply connected which is expressed as $\pi_1(S^n : n \geq 2) = 0$. On the other hand, the circle S^1 is infinitely connected, namely, the number of homotopic classes for the paths is infinite as is shown later in Figure D.3.

We see that relations $R(\theta+\pi)=R(\theta-\pi)=-R(\theta)$ and $R(\theta+2\pi)=R(\theta)$ hold for $SO(3)$, but in $SU(2)$, $U(\theta)=-U(\theta+2\pi)$ holds. One can easily see there is one-to-two correspondence between $SU(2)$ and $SO(3)$ by comparing, for instance, a rotation on the z-axis by setting $\mathbf{n}=(0,0,1)$ in Eq. (D.1) and $\theta=\beta=0$ in Eq. (D.3). It is the origin of the well known fact that a spinor field changes its sign by 2π rotation but a vector does not. Since a rotation in three-dimensional real space can be parameterized by specifying the axis of rotation and rotation angle, its manifold can be defined as the space spanned by the rotation vectors whose length and direction equal to the rotation angle and rotation axis. It is a sphere (R^3) with radius π. However, the π rotation and $-\pi$ rotation is the same, which means that diametric points on the surface should be identified. The manifold is said to be doubly connected. This is because two lines can be envisaged to represent a series of continuous rotation (see Figure D.2). The path denoted as g_1 is a series of rotation whose rotation angle is smaller than π. The path can be converged to a point and belongs to the same homotopy class as a point. On the other hand the path g_2 is a case for $\theta>\pi$. When it reaches π at A and goes beyond it, it jumps to A′ on the other side. When it makes a closed loop by going back to where it started, the loop g_2 cannot converge to a point. Therefore it belongs to a different homotopy class. When the rotation angle exceeds 2π it jumps again to opposite side and returns to where it started, namely to the same homotopy class as a point. If one makes a correspondence $g_1\to 0$, $g_2\to 1$, then they constitute a two-integer group referred to as Z_2 which satisfies algebra $0+0=0$, $0+1=1+0=1$, $1+1=0$. In mathematical language it is written as

$$\pi_1(SO(3))=Z_2 \tag{D.4}$$

Both $SO(3)$ group and $SU(2)$ group satisfy the same Lie algebra and both are compact. However, $SU(2)$ is simply connected while $SO(3)$ is doubly connected.

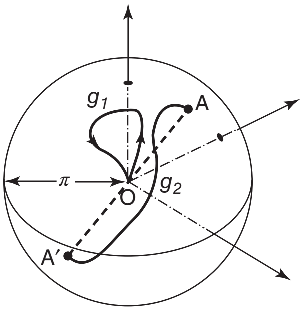


Figure D.2 The rotation in three-dimensional space can be parameterized by its direction of rotation axis and rotation angle. A space spanned by vectors with length equal to the rotation angle pointing to the rotation axis is a three-dimensional sphere (R^3) with radius π. As $+\pi$ rotation and $-\pi$ rotation is identical, all the points on the surface can be identified with those on the opposite side. A series of continuous rotation makes a path within the sphere. The path g_1 can be shrank to a point in the origin (no rotation) but g_2 whose rotation angle exceeds π cannot deform to overlap with g_1 by a continuous transformation.

In this case, $SU(2)$ is referred to as *the covering group of the algebra*. The $SO(3)$ can be obtained from $SU(2)$ by identifying Z_2, (the subgroup of $SU(2)$ consisting of two elements, 1 and $e^{i\pi}$) with identity transformation. It can be expressed concisely as

$$SO(3) = SU(2)/Z_2 \tag{D.5}$$

The $SO(3)$ is said to be a coset of $SU(2)$ (see discussions on page 557). In general, the group $SU(N)$ contains a subgroup $\exp[i2\pi/n]$ $(n = 0, 1, \cdots, N-1)$ which commutes with all other elements in the group.

Degenerate vacua and the winding number: A necessary condition for a soliton solution to exist in the static equation of motion for a scalar field $\phi(\mathbf{r})$ is that degenerate multi-vacua with different homotopy number exist and that it has finite energy localized in space. Given a potential $V(\phi)$, the degenerate vacua appear at $\phi = v$ when the field ϕ satisfying conditions

$$V(\phi = v) = 0, \quad \left.\frac{\partial V}{\partial \phi}\right|_{\phi=v} = 0, \quad \left.\frac{\partial^2 V}{\partial \phi^2}\right|_{\phi=v} > 0 \qquad r = |\mathbf{r}| \to \infty \tag{D.6}$$

is multi-valued at $\phi = v$. In other words, given a field ϕ and the equation of motion having an internal symmetry G, the first step to to find a soliton solution is to look for a condition that

$$\phi(r \to \infty) = v \tag{D.7}$$

is multi-valued. As the space point $r = |\mathbf{r}|$ has the topology of S^1 in two-dimension and S^2 in three-dimension at $r \to \infty$, we need to know which of the corresponding manifold is picked up by the field $\phi(\mathbf{r})$, that is, by the projection $\mathbf{r} \to \phi \subset G$.

Two-dimensional space: We consider a Higgs field $\varphi(\mathbf{r}) = (\phi_1 + i\phi_2)/\sqrt{2}$ in two-dimensional space of ϕ_i that satisfies a $U(1)$ internal symmetry. Let us assume that the vacuum configuration satisfies

$$\begin{aligned} |\varphi(r \to \infty)| = \sqrt{|\phi_1(\mathbf{r})|^2 + |\phi_2(\mathbf{r})|^2} = v\exp[-ig(\theta)], \\ g(2\pi) = g(0) + 2\pi n \ (n = 0,\ 1,\ 2, \cdots) \end{aligned} \tag{D.8}$$

The variable $r = \sqrt{x^2 + y^2} \to \infty$ has the same topology as a circle S^1 and can be parameterized by an angle variable θ where $\theta = 2\pi$ is identified with $\theta = 0$. According to considerations we gave above, the field $\varphi(\mathbf{r})$ represents a projection from a point $\mathbf{r}$ in real space to φ in the φ space. Let us consider the following group variables (see Figure D.3).

$$g_1(\theta) = 0 \qquad 0 \le \theta < 2\pi \tag{D.9a}$$

$$g_2(\theta) = \begin{cases} \alpha\theta & 0 \le \theta < \pi \\ \alpha(2\pi - \theta) & \pi \le \theta < 2\pi \end{cases} \tag{D.9b}$$

$$g_3(\theta) = \theta \qquad 0 \le \theta < 2\pi \tag{D.9c}$$

$$g_4(\theta) = 2\theta \qquad 0 \le \theta < 2\pi \tag{D.9d}$$

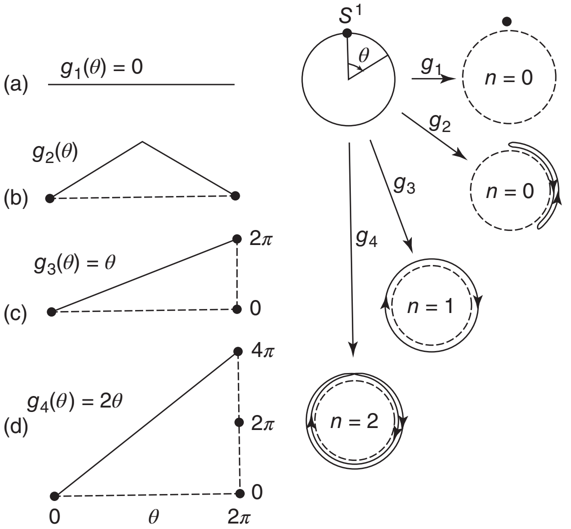


Figure D.3 Homotopy classification of the projection $S^1 \to S^1$. It means that expressing a point at infinity in two-dimensional space of the spacetime as $S^1 (0 \leq \theta < 2\pi)$, its projection on to $g(\theta)$ in the g (group) space represents also S^1, namely a one-dimensional loop or a straight line with both ends identified.

$g_1(\theta)$ represents identity transformation. Obviously, $g_2(\theta)$ can be brought to the same $g_1(\theta)$ by continuous transformations hence belongs to the same homotopy class as g_1. However, g_3 winds once around the circle S^1 and cannot be transformed to g_1 by continuous transformations. Similarly, g_4 winds twice around the circle and belongs to a different homotopy class from g_1 and g_3. We define a topological index or the winding number by

$$n = \frac{1}{2\pi} \oint_0^{2\pi} dg = \frac{i}{V_G} \oint U^{-1} \frac{\partial U}{\partial \theta} d\theta \tag{D.10}$$

V_G is volume of the group manifold which is 2π in the present case. The integrand of the last equality is an operation to extract the group variable from the group element U. When θ varies from $0 \to 2\pi$, the group variable g changes from $0 \to 2n\pi$. Since $n(g_1) = n(g_2) = 0$, $n(g_3) = 1$, $n(g_4) = 2$, there exist multi-vacua distinguished by an integer n. When $S^1 \to S^1$ projection can be classified to homotopy classes distinguished by the homotopy index n, it is expressed as

$$\pi_1(S^1) = Z \tag{D.11}$$

and is referred to as the *homotopy group* of the first kind. An ensemble of integers with operation of addition and subtraction make an integer group Z. When the group in question is expressed explicitly, it is also written as

$$\pi_1(U(1)) = Z \tag{D.12}$$

In general, homotopy classifications for the projections $S^n \to S^m$, $S^n \to G$ are expressed as $\pi_n(S^m)$, $\pi_n(G)$.

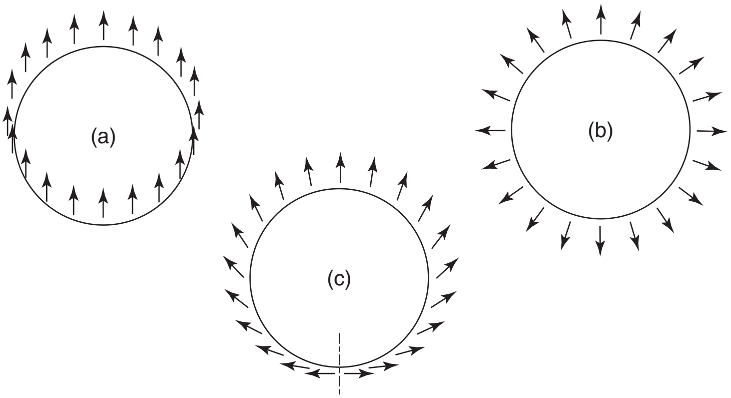


Figure D.4 $\hat{\boldsymbol{\phi}}(x, y, z)$ is a projection from a three-dimensional space point to another three-dimensional point in the $\hat{\boldsymbol{\phi}}$ space. If $\hat{\phi}_1^2 + \hat{\phi}_2^2 + \hat{\phi}_3^2 = 1$ for $r \to \infty$, it is a projection of surface of a sphere ($S^2 : r = \sqrt{x^2 + y^2 + z^2} \to \infty$) in the three-dimensional space of (x, y, z) to $\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3)$ which is a unit vector ($\hat{\boldsymbol{\phi}} \cdot \hat{\boldsymbol{\phi}} = 1$) in the isospin space. (a) $\hat{\boldsymbol{\phi}} = (0, 0, 1)$, (b) $\hat{\boldsymbol{\phi}} = (\hat{\xi}, \hat{\eta}, \hat{\zeta})$, $|\hat{\boldsymbol{\phi}}| = \sqrt{\hat{\xi}^2 + \hat{\eta}^2 + \hat{\zeta}^2} = 1$. Isospin vectors on the surface of a sphere in (b) cannot be converted continuously to those in (a) at all points on the surface. (c) By punching a hole at the south pole and excluding the hole point, the continuous transformation from (b) to (a) is possible.

Three-dimensional space: Let us consider an isospin field $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)$ that satisfies the $SO(3)$ symmetry, that is,

$$\begin{aligned} \hat{\boldsymbol{\phi}} \cdot \hat{\boldsymbol{\phi}} &= \hat{\phi}^2 = \hat{\phi}_1^2 + \hat{\phi}_2^2 + \hat{\phi}_3^2 = 1 \qquad \text{for } r = |\boldsymbol{r}| \to \infty \\ \hat{\phi} &= \frac{\phi}{|\boldsymbol{\phi}(r \to \infty)|} = \frac{\phi}{v} \end{aligned} \tag{D.13}$$

The field $\hat{\boldsymbol{\phi}}$ is a unit vector which directs in any direction in the isospin space. We choose the vacuum expectation value as $\hat{\phi} = (0, 0, 1)$. After the spontaneous symmetry breakdown from $SO(3) \to U(1)$, the vacuum is still invariant by rotations on the axis in the $\hat{\boldsymbol{\phi}} = (0, 0, 1)$ direction [$U(1)$ or $SO(2)$]. As the configuration of $\hat{\boldsymbol{\phi}}$ represents a surface of unit sphere, $\hat{\phi}_a(r \to \infty)$ represents the projection from S^2 in the physical space to S^2 in the isospin space and is expressed as $\pi_2(S^2)$. From considerations of the topology, it is known [947, 948] that

$$\pi_2(S^2) = Z^{2)} \tag{D.14}$$

An intuitive illustration of Eq. (D.14) is given in Figure D.4. Figure D.4a illustrate vectors $\hat{\boldsymbol{\phi}} = (0, 0, 1)$, (b) $\hat{\boldsymbol{\phi}} = (\xi, \eta, \zeta)$, ($\sqrt{\xi^2 + \eta^2 + \zeta^2} = 1$). It is not possible to change the orientation of $\hat{\boldsymbol{\phi}}$ in the configuration (Figure D.4b) continuously to those in (Figure D.4a) at all points. However if one allows an exception at a point, for instance discontinuity at the south pole, it is possible to bring $\hat{\boldsymbol{\phi}}$ in (Figure D.4b) to that in (Figure D.4a) by continuous transformation like one in (Figure D.4c). In

---

2) In general, $\pi_n(S^n) = Z$, $\pi_n(S^m) = 0$ $(n < m)$.

other words, the configuration in (Figure D.4a) and that in (Figure D.4b) belong to different homotopy classes. The index to specify the homotopy class can be defined by

$$Q = \frac{1}{8\pi}\int d^3x\, \epsilon_{ijk}\epsilon^{abc}\partial_i\hat{\phi}^a\partial_j\hat{\phi}^b\partial_k\hat{\phi}^c \tag{D.15}$$

where Q is referred to as the *winding number* or *Pontryagin charge*. To see that Q is an integer, we use the totally antisymmetric nature of ϵ_{ijk} to rewrite

$$\partial_i\hat{\phi}^a\partial_j\hat{\phi}^b\partial_k\hat{\phi}^c = \partial_i(\hat{\phi}^a\partial_j\hat{\phi}^b\partial_k\hat{\phi}^c) \tag{D.16}$$

and convert Eq. (D.15) to a surface integral to obtain

$$Q = \frac{1}{8\pi}\int dS_i\, \epsilon_{ijk}\epsilon^{abc}\hat{\phi}^a\partial_j\hat{\phi}^b\partial_k\hat{\phi}^c \tag{D.17}$$

Next we change integration variables in Eq. (D.17) to angular variables θ, α. The infinitesimal surface element can be expressed as

$$dS_i = \epsilon_{imn}\frac{\partial x_m}{\partial\theta}\frac{\partial x_n}{\partial\alpha}d\theta d\alpha \tag{D.18}$$

Then

$$\begin{aligned} Q &= \frac{1}{8\pi}\int d\theta d\alpha\, \epsilon_{ijk}\epsilon^{abc}\hat{\phi}^a\partial_j\hat{\phi}^b\partial_k\hat{\phi}^c\epsilon_{imn}\frac{\partial x_m}{\partial\theta}\frac{\partial x_n}{\partial\alpha} \\ &= \frac{1}{4\pi}\int d\theta d\alpha\epsilon^{abc}\hat{\phi}^a\frac{\partial(\hat{\phi}^b,\hat{\phi}^c)}{\partial(\theta,\phi)} \end{aligned} \tag{D.19}$$

Equation (D.19) to define the Q is an operation to extract the group variable similar to that of $U(1)$ in Eq. (D.10). Among the factors in the integrand, those to the right of $d\theta d\alpha$ is a Jacobian to convert variables $\hat{\boldsymbol{\phi}} \to (\theta, \alpha)$ hence Eq. (D.19) can be reexpressed as an integral in the variable $\hat{\boldsymbol{\phi}}$.

$$Q = \frac{1}{4\pi}\int dS_\phi^a\hat{\phi}^a = \frac{1}{4\pi}\int dS_\phi \tag{D.20}$$

As S^2 in the $\hat{\boldsymbol{\phi}}$ space is a surface of a unit sphere, if x_i sweeps the whole surface of S^2 in spacetime, Q gives the number n of how many times the $\hat{\boldsymbol{\phi}}$ covers the surface of the sphere in $\hat{\boldsymbol{\phi}}$ space. The configuration in Figure D.4a corresponds to $n = 0$ and that in (b) to $n = 1$.

Coset group: The operation of $SO(3)$ can be defined by specifying the rotation axis and the rotation angle. The vacuum configuration we discussed above constrains (fixes) the rotation axis but not the angle. The rotation on the fixed axis [$U(1)$ operation] keeps the vacuum invariant, hence there are as many freedoms to define the vacuum as rotations on the axis. When one extracts group H from the original group G, and writing elements that do not belong to H as M_i, the group G can be written formally as

$$G = H \oplus HM_1 \oplus HM_2 \oplus \cdots \tag{D.21}$$

The space spanned by M_i's is referred to as *coset* and is denoted as G/H. From the above discussion, one sees that, topologically, the coset $SO(3)/U(1)$ is homomorphic3) to S^2.

---

3) If the two groups G and G' have elements $g \subset G$, $g' \subset G'$ that can be put into correspondence to each other when their multiplication tables are compared, the two groups are said to be homomorphic. If the correspondence is one-to-one, the group is isomorphic.

Appendix E
General Relativity

E.1
Geodesic Equation

Geodesic equation as the parallel transport of a tangential vector: Let us first, heuristically, introduce the "connection", which by definition parallel-transports a spacetime vector $p^\mu(x)$ located at a point x to a nearby point $x + dx$.[1]

$$\begin{aligned} p^\mu(x + dx^\nu)_\parallel &= p^\mu(x) - \Gamma^\mu{}_{\lambda\nu} p^\lambda dx^\nu, \\ p_\mu(x + dx^\nu)_\parallel &= p_\mu(x) + \Gamma^\lambda{}_{\mu\nu} p_\lambda dx^\nu. \end{aligned} \tag{E.1}$$

Parallel transport means that the vector is moved without changing its length and direction. Then, the covariant derivative, which is covariant under general coordinate transformations, is obtained from Eq. (E.1).

$$\begin{aligned} D_\nu p^\mu &= \partial_\nu p^\mu + \Gamma^\mu{}_{\lambda\nu} p^\lambda \equiv p^\mu{}_{;\nu}, \\ D_\nu p_\mu &= \partial_\nu p_\mu - \Gamma^\lambda{}_{\mu\nu} p_\lambda \equiv p_{\mu;\nu}, \\ D_\lambda T^\mu{}_\nu &= \partial_\lambda T^\mu{}_\nu + \Gamma^\mu{}_{\sigma\lambda} T^\sigma{}_\nu - \Gamma^\sigma{}_{\lambda\nu} T^\mu{}_\sigma. \end{aligned} \tag{E.2}$$

The covariant derivative of the tensor can be derived by considering that of the product of two vectors.

The Einstein's equivalence principle states that the motion of a particle in the gravitational field becomes the free-fall equation in the Minkowsky space. The free-fall (law of inertia) equation in the general coordinate system can be obtained by using the covariant derivative. In other words, it is obtained by dividing the momentum vector by the proper time τ, and becomes

$$\frac{Dp^\mu}{D\tau} = 0 \quad \to \quad \frac{dp^\mu}{d\tau} + \Gamma^\mu{}_{\lambda\nu} p^\lambda \frac{dx^\nu}{d\tau} = 0. \tag{E.3}$$

The conventional form for the geodesic equation is obtained by dividing with mass m.

$$\frac{d^2 x^\mu}{d\tau^2} + \Gamma^\mu{}_{\lambda\nu} \frac{dx^\lambda}{d\tau} \frac{dx^\nu}{d\tau} = 0 \tag{E.4}$$

---

[1] We do not derive the general relativistic equations explicitly. It is beyond the scope of the book. The interested readers may refer to the standard textbooks of general relativity. For equations in the Section E.3 and E.4, the reader may refer to [125, 787, 835, 949]

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

The expression for the connection $\Gamma^{\lambda}{}_{\mu\nu}$ for the general coordinate transformations is referred to as the *Christoffel symbol* and is given by

$$\Gamma^{\sigma}{}_{\rho\nu} \equiv g^{\sigma\lambda}\Gamma_{\lambda\rho\nu} = \frac{1}{2}g^{\sigma\lambda}\left(\partial_\rho g_{\nu\lambda} + \partial_\nu g_{\lambda\rho} - \partial_\lambda g_{\rho\nu}\right) = g^{\sigma\lambda}\eta_{\alpha\beta}\frac{\partial\xi^{\alpha}}{\partial x^{\lambda}}\frac{\partial^2\xi^{\beta}}{\partial x^{\rho}\partial x^{\nu}}. \tag{E.5}$$

where ξ^{α}, ξ^{β} are spacetime coordinates in the Minkowski space.

The Newtonian limit: In the nonrelativistic approximation ($v \ll 1$) and weak static gravitational potential, we show that the geodesic equation becomes the Newtonian equation

$$m\frac{d^2\boldsymbol{x}}{dt} = -\nabla(m\Psi) \tag{E.6}$$

where Ψ is the Newtonian gravitational potential. As the velocity is small, $d\boldsymbol{x} \ll dt$, the dominant component of Eq. (E.4) is

$$\frac{d^2x^{\mu}}{d\tau^2} + \Gamma^{\mu}{}_{00}\left(\frac{dt}{d\tau}\right)^2 = 0. \tag{E.7}$$

In the approximation, the metric tensor is a small deviation from the Minkowski metric tensor. Consequently, we can write

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \tag{E.8}$$

where $\eta_{\mu\nu} = \text{dia}(1, -1, -1, -1)$. Using the expression for the Christoffel connection

$$\Gamma^{\mu}{}_{\lambda\nu} = \frac{g^{\mu\rho}}{2}\left(\partial_\nu g_{\lambda\rho} + \partial_\lambda g_{\nu\rho} - \partial_\rho g_{\lambda\nu}\right) \tag{E.9}$$

we obtain to the first order in $h_{\mu\nu}$

$$\Gamma^{\mu}{}_{00} \simeq \frac{\eta^{\mu\rho}}{2}\frac{\partial h_{00}}{\partial x^{\rho}}. \tag{E.10}$$

Then the geodesic equation becomes

$$\frac{d^2\boldsymbol{x}}{d\tau^2} = -\frac{1}{2}\left(\frac{dt}{d\tau}\right)^2\nabla h_{00}, \tag{E.11}$$

$$\frac{d^2t}{d\tau^2} = 0. \tag{E.12}$$

From Eq. (E.12), we have $dt/d\tau =$ constant. Dividing both sides of Eq. (E.11) by $(dt/d\tau)^2$, we obtain

$$\frac{d^2\boldsymbol{x}}{dt^2} = -\frac{1}{2}\nabla h_{00}. \tag{E.13}$$

Comparing Eq. (E.13) with Eq. (E.6), we obtain

$$h_{00} = 2\Psi + \text{constant}. \tag{E.14}$$

At far distance from the mass that produces the potential, the metric should be Newtonian (i.e., Minkowskian), hence the constant should be set to zero. The metric is thus

$$g_{00} = 1 + 2\Psi. \tag{E.15}$$

E.2
Ricci Tensor and Scalar

The Riemann tensor is defined by

$$R^{\mu}{}_{\nu\alpha\beta} = \partial_{\alpha}\Gamma^{\mu}{}_{\nu\beta} - \partial_{\beta}\Gamma^{\mu}{}_{\nu\alpha} + \Gamma^{\mu}{}_{\sigma\alpha}\Gamma^{\sigma}{}_{\nu\beta} - \Gamma^{\mu}{}_{\sigma\beta}\Gamma^{\sigma}{}_{\nu\alpha}. \tag{E.16}$$

The Ricci tensor and Ricci scalar are defined by

$$R_{\mu\nu} = R^{\lambda}{}_{\mu\lambda\nu}, \qquad \mathscr{R} = g^{\mu\nu}R_{\mu\nu}.^{2)} \tag{E.17}$$

The Einstein equation of motion is given by

$$\mathscr{G}_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}\mathscr{R} = 8\pi G_N T_{\mu\nu}. \tag{E.18}$$

where $T_{\mu\nu}$ is the energy-momentum tensor.

An observer at rest in the comoving coordinate system remains at rest, that is, (r, θ, ϕ) remain unchanged. If one introduces a homogeneous, isotropic fluid initially at rest in this frame, the t =constant hypersurfaces will always be orthogonal to the fluid flow and will always coincide with the hypersurfaces having both spatial homogeneity and constant fluid density. Starting from the spatial part of the metric,

$$d\boldsymbol{l}^2 = h_{ij}dx^i dx^j \tag{E.19}$$

the three-dimensional Riemann tensor, Ricci tensor, and Ricci scalar can be defined similarly as Eq. (E.16) and (E.17) [125, 950].

$$\begin{aligned} R^{(3)l}{}_{ijk} &= \partial_j\Gamma^{l}{}_{ik} - \partial_k\Gamma^{l}{}_{ij} + \Gamma^{l}{}_{mj}\Gamma^{m}{}_{ik} - \Gamma^{l}{}_{mk}\Gamma^{m}{}_{ij}, \\ R^{(3)}_{ij} &= R^{(3)l}{}_{ilj}, \\ \mathscr{R}^{(3)} &= h^{ij}R^{(3)}_{ij}. \end{aligned} \tag{E.20}$$

E.3
Gauge Degrees of Freedom

A perturbed flat Robertson–Walker metric can be written in general as

$$\begin{aligned} g_{00} &= \left[1 + 2A(t, \boldsymbol{x})\right], \\ g_{0i} &= R(t)\, w_i(t, \boldsymbol{x}), \\ g_{ij} &= -R^2(t)\left\{[1 + 2C(t, \boldsymbol{x})]\gamma_{ij} + h_{ij}(t, \boldsymbol{x})\right\}, \qquad \gamma_{ij}h_{ij} = 0 \end{aligned} \tag{E.21}$$

where the functions A, w_i, C, h_{ij} represent the metric perturbations about the Robertson–Walker spacetime and assumed to be small compared to unity. In all practical applications, the first-order perturbation is known to give a good approximation (linear regime). The trace part $(\sum \gamma_{ij}h_{ij})$ of the perturbation to g_{ij} is absorbed in C and h_{ij} is taken to be traceless. In the following, we simplify

---

2) We use $\mathscr{R}$ for the scalar curvature to avoid confusion with the cosmic scale factor $R(t)$.

arguments by assuming a flat background, i.e., unperturbed universe for which $\gamma_{ij} = \delta_{ij}$ for Cartesian coordinates.

There are 10 independent g_{ij}s, which translate to two scalars A, C, one vector w_i with three independent components, and one tensor h_{ij} with five independent components. Note they are a vector and a tensor in three-dimensional space. Actually, two more scalars can be defined. The longitudinal component of the vector can be expressed as $\nabla_i B(t, \boldsymbol{x})$, which is separated from the transverse (or rotational) component and can be classified as a scalar.[3] Similarly, those parts of the tensor that can be expressed as $\nabla_i \nabla_j E(t, \boldsymbol{x})$ (scalar components) and those that can be expressed as $\nabla_i E_j$ (vector components) can also be separated. As a result, we have four scalars, two vectors with two independent components, and a tensor with two independent components. However, not all of them are physical. Translational coordinate transformations $x^\mu \to \tilde{x}^\mu + \xi(t, \boldsymbol{x})$ do not change the physical content, and hence are left as the freedom of gauge transformations. Therefore, out of the 10 degrees of freedom, only 6 have physical meaning: 2 for scalars, 2 for the vector, and 2 for the tensor. Among these, the tensor perturbation is already gauge invariant. Similar decomposition can be made for the Einstein tensor ($G_{\mu\nu} = R_{\mu\nu} - g_{\mu\nu}\mathscr{R}/2$) and the energy–momentum tensor $T_{\mu\nu}$. In the linear regime, Fourier components of the three modes (scalar, vector, and tensor) do not mix and their dynamical evolutions can be solved independent of the others. The rotational part of the vector metric has only time-decreasing solutions (see Eq. (9.28a) in the text) and the tensor represents a gravitational wave.

Using this gauge freedom, one can modify the metric in a form convenient for solving the equations one wants to treat. Traditionally, the synchronous gauge has been used, but recently the conformal Newtonian gauge (also known as *longitudinal gauge*) has become more conventional.

From now on, we restrict ourselves to the scalar perturbation only. Then Eqs. (E.21) for a flat universe are reduced to

$$\begin{aligned} g_{00} &= \left[1 + 2\Psi(t, \boldsymbol{x})\right], \\ g_{0i} &= R(t)\, \nabla_i B(t, \boldsymbol{x}), \\ g_{ij} &= -R^2(t)\left[\left\{1 + 2\Phi(t, \boldsymbol{x})\right\}\delta_{ij} - 2\nabla_i \nabla_j E\right]. \end{aligned} \tag{E.23}$$

Two gauge-invariant scalar variables can be defined as [951]

$$\begin{aligned} \Phi_A &\equiv \Psi - \frac{\partial}{\partial t}\left[R\left(B - R\dot{E}\right)\right], \\ \Phi_H &\equiv -\Phi + RH\left(B - R\dot{E}\right). \end{aligned} \tag{E.24}$$

where the dot means time derivative. By choosing B, E suitably, we can obtain the synchronous and conformal Newtonian gauges.

---

[3] In the Fourier space

$$\begin{aligned} \boldsymbol{v} &\equiv \boldsymbol{v}_{\parallel} + \boldsymbol{v}_{\perp} = \frac{\boldsymbol{k}(\boldsymbol{k}\cdot\boldsymbol{v})}{|\boldsymbol{k}|^2} + \left\{\boldsymbol{v} - \frac{\boldsymbol{k}(\boldsymbol{k}\cdot\boldsymbol{v})}{|\boldsymbol{k}|^2}\right\}, \\ &\boldsymbol{k}\cdot\boldsymbol{v}_{\perp} = 0, \qquad \boldsymbol{k}\times\boldsymbol{v}_{\parallel} = 0. \end{aligned} \tag{E.22}$$

Synchronous gauge: In the synchronous gauge, one sets $g_{00} = g_{0i} = 0$, or $\Psi = B = 0$ in Eq. (E.23). It was pioneered by Lifshitz [952, 953], and has been used in many standard textbooks [954, 955]. However, it has some singularities. It also has redundant freedom of gauge transformation, which aroused some confusions. The confusion can be avoided by using gauge-invariant variables.

Conformal Newtonian gauge: The conformal Newtonian gauge can be obtained by setting $B = E = 0$ [956, 957]. One gets the metric

$$ds^2 = \left[(1 + 2\Psi)dt^2 - R^2(t)(1 + 2\Phi)dx^i dx_i\right] .^{4)} \tag{E.25}$$

Metrics of the conformal Newtonian gauge are directly connected to the Newtonian mechanics. The Ψ in the time component of the metric tensor is the Newton's gravitational potential, as was already shown in Eq. (E.15). Φ measures the spatial curvature of constant-density hypersurfaces $\mathscr{R}^{(3)} = -4\nabla^2\Phi/R^2$, which can easily be derived by substituting the metric in Eq. (E.20).

It can be shown working through the Einstein equation that $\Phi \approx -\Psi$. The difference is induced by the anisotropic stress components of radiations, which are generally small. For perfect fluids, which is the assumption in the main text, $\Phi = -\Psi$ holds.

E.4
Gravitational Waves

We saw that, in the linear regime, the traceless tensor component is decoupled from the scalar and vector parts and can be treated completely independently. It has only two independent components, and in the orthogonal metric where $\gamma_{ij} = \delta_{ij}$, the tensor part of the metric can be expressed as

$$g_{ij} = R^2 \begin{bmatrix} 1 + h_+ & h_\times & 0 \\ h_\times & 1 - h_+ & 0 \\ 0 & 0 & 1 \end{bmatrix} . \tag{E.26}$$

[5)]The tracelessness means $k^j h_{ij} = 0$; hence by adopting the above metric we have implicitly chosen a coordinate system in which the tensor perturbation has no z (longitudinal) component. The action for the gravity is given by (see Eq. (6.81))

$$\begin{aligned} S &= \int d^4x \sqrt{-g}\,(\mathscr{L}_g + \mathscr{L}_m) \\ \mathscr{L}_g &= \frac{\mathscr{R}}{16\pi G_N} = \frac{\tilde{M}_{\text{Pl}}^2}{2}\mathscr{R},^{5)} \qquad g = \det g_{\mu\nu} . \end{aligned} \tag{E.27}$$

---

4) Nomenclature for the scalar metric in the Newtonian gauge is very confusing. We follow that of Dodelson [787]. In Peacock's book [949], $\Psi \to \phi$, $\Phi \to -\psi$; in [958] $\Psi \to \Psi$, $\Phi \to -\Phi$; in [835] $\Psi \to \Phi$, $\Phi \to \Psi$.

5) $\tilde{M}_{\text{Pl}}$ is denoted as $\overline{M}_{\text{Pl}}$ in Chapter 6.

Working out the Einstein equation, one sees that the density component couples only to the scalar part and does not couple to the tensor component [787, 958] (see also Eq. (6.99)). The end result for the equation of motion becomes

$$\ddot{h}_a + 3H\dot{h}_a + k^2 h_a = 0 \tag{E.28}$$

where $a = +, \times$. Equation (E.28) is the same wave equation as the massless scalar field. As the action has a normalization factor $1/16\pi G_N = \tilde{M}_{\text{Pl}}^2/2$, the effective scalar field is given by $\phi_a = \tilde{M}_{\text{Pl}} h_a/\sqrt{2}$ or $h_a = \sqrt{2}\phi_a/\tilde{M}_{\text{Pl}}$; hence the amplitude of the tensor components is suppressed by the Planck mass. The corresponding solutions are called *gravitational waves*. In the limit $k \to 0$, it has a constant solution. That is, the super-horizon wave has a constant amplitude. For $kH \gtrsim 1$, it is a damped oscillation. It means that the sub-horizon component of the gravity wave decays once it enters the horizon. Therefore, considering the effect of gravity waves on CMB anisotropy spectrum, small-scale components can be neglected. Only the large-scale anisotropies are impacted by the gravity wave.

Appendix F
Tensor Spherical Harmonic Function

Properties of the generalized spin spherical harmonic Functions [794, 796]: We denote the spherical harmonic functions having spin s ($s \neq 0$) as ${}_sY^*_{lm}(\theta,\phi)$. These sets of functions (one for each particular spin) satisfy the same completeness and orthonormality relations as the standard spherical harmonic functions, that is

$$\begin{aligned}
&\int d\Omega\, {}_sY^*_{l',m'}(\theta,\phi)\, {}_sY_{lm}(\theta,\phi) = \delta_{l'l}\delta_{m'm},\\
&\sum_{lm} {}_sY^*_{lm}(\theta',\phi')\, {}_sY_{lm}(\theta,\phi) = \delta(\phi-\phi')\delta(\cos\theta-\cos\theta').
\end{aligned} \tag{F.1}$$

${}_sY^*_{lm}(\theta,\phi)$ ($s \neq 0$) is related to the standard Y_{lm} by the relation

$$\begin{aligned}
{}_sY_{lm} &= \left[\frac{(l-s)!}{(l+s)!}\right]^{\frac{1}{2}} \eth^s Y_{lm}, \qquad (0 \le s \le l),\\
{}_sY_{lm} &= (-1)^s \left[\frac{(l+s)!}{(l-s)!}\right]^{\frac{1}{2}} \bar{\eth}^{-s} Y_{lm}, \qquad (-l \le s \le 0)
\end{aligned} \tag{F.2}$$

where $\eth$ and $\bar{\eth}$ are the spin-raising and spin-lowering operators, respectively, defined by

$$\begin{aligned}
\eth\, {}_sf(\theta,\phi) &= -\sin^s(\theta)\left[\frac{\partial}{\partial\theta} + i\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\right]\sin^{-s}\theta\, {}_sf(\theta,\phi),\\
\bar{\eth}\, {}_sf(\theta,\phi) &= -\sin^{-s}(\theta)\left[\frac{\partial}{\partial\theta} - i\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\right]\sin^{s}\theta\, {}_sf(\theta,\phi).
\end{aligned} \tag{F.3}$$

Here, ${}_sf(\theta,\phi)$ is a function defined on a sphere. The spherical harmonic functions also satisfy relations

$$\begin{aligned}
{}_sY^*_{lm} &= (-1)^s\, {}_{-s}Y_{l,-m},\\
\eth\, {}_sY_{lm} &= [(l-s)(l+s+1)]^{1/2}\, {}_{s+1}Y_{l.m},\\
\bar{\eth}\, {}_sY_{lm} &= -[(l+s)(l-s+1)]^{1/2}\, {}_{s-1}Y_{l.m},\\
\bar{\eth}\,\eth\, {}_sY_{lm} &= -(l-s)(l+s+1)\, {}_sY_{lm}.
\end{aligned} \tag{F.4}$$

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

Appendix G
Destiny of the Cosmos

We consider the dynamical evolution of the Universe under various conditions. Let us express explicitly the scale dependence of energy densities of matter ($\rho_m \sim R^{-3}$), radiation ($\rho_{\text{rad}} \sim R^{-4}$), and vacuum ($\rho_v \sim$ constant). Defining

$$\begin{aligned} \Omega_m &= \frac{\rho_m(t_0)}{\rho_c}, \qquad \Omega_{\text{rad}} = \frac{\rho_{\text{rad}}(t_0)}{\rho_c}, \qquad \rho_c = \frac{3H_0^2}{8\pi G_N}, \\ K &= -\frac{k}{R_0^2 H_0^2}, \qquad \Omega_\Lambda = \frac{\Lambda}{3H_0^2} = \frac{\rho_v}{\rho_c} \end{aligned} \tag{G.1}$$

we modify the Friedman equation

$$H^2 = \left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G_N}{3}(\rho_m + \rho_{\text{rad}} + \rho_v) - \frac{k}{R^2} \tag{G.2}$$

to

$$\begin{aligned} &\frac{1}{R_0^2 H_0^2}\left(\frac{dR}{dt}\right)^2 + U(R) = K \\ &U(R) = -\left[\Omega_m\left(\frac{R_0}{R}\right) + \Omega_{\text{rad}}\left(\frac{R_0}{R}\right)^2 + \Omega_\Lambda\left(\frac{R}{R_0}\right)^2\right]. \end{aligned} \tag{G.3}$$

Introducing $d\tau = H_0 dt$, $x = R/R_0$, the behavior of the Universe is reduced to a one-dimensional mass point motion in the potential

$$\begin{aligned} &\left(\frac{dx}{dt}\right)^2 + U(x) = K, \qquad U(x) = -\frac{\Omega_m}{x} - \frac{\Omega_{\text{rad}}}{x^2} - \Omega_\Lambda x^2, \\ &K = 1 - \Omega_m - \Omega_{rad} - \Omega_\Lambda. \end{aligned} \tag{G.4}$$

Case (1) $\Omega_\Lambda < 0$:
The potential shape is depicted as the dashed curve in Figure G.1(a). The potential is positive and $U(x) \to \infty$ as $x \to \infty$; if the universe starts at $x = 0$, it is bounced back by the potential barrier and contracts again.

Case (2) $\Omega_\Lambda = 0$:
As $U(x \to \infty) \to 0$, the Universe's fate is recontraction or eternal expansion depending on whether $K < 0$ or ≥ 0. Because of Eq. (G.3), there is one-to-

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

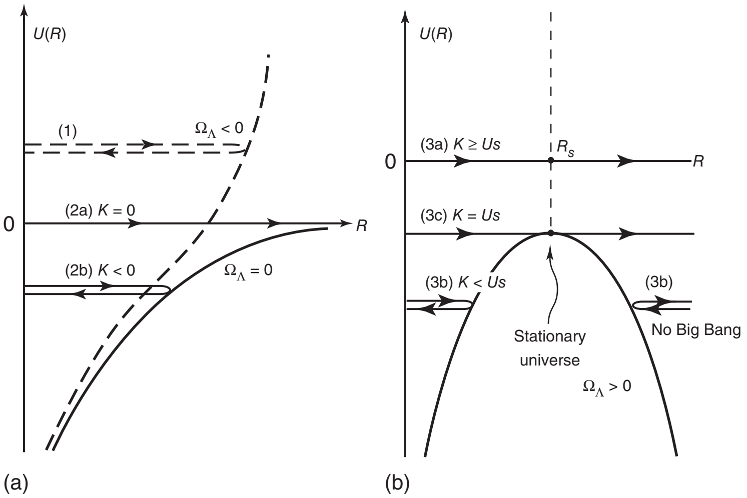


Figure G.1 Cosmic destiny (a) Case (1) $\Omega_\Lambda < 0$: expansion/contraction. Case (2) $\Omega_\Lambda = 0$. Depending on $K \geq 0$ (2a) or $K < 0$ (2b), it means eternal expansion or recontraction. Note $K = -k/R_0^2 H_0^2$. (b) Case (3) $\Omega_\Lambda > 0$: Depending on the relative size of $U_s = \max[U(R)]$ and K, several branches exist. See text.

one correspondence between the critical density $\rho_c \gtreqless 1$ and the curvature $k \gtreqless 0$.

Case (3) $\Omega_\Lambda > 0$:
In this case, $U(x \to \infty) \to -\infty$: Three possibilities exist depending on whether $K \gtreqless U_s \equiv \max[U(x)]$ (see Figure G.1(b)).

Case (3a) $K > U_s$: Here, the expansion rate is always positive $\dot{x} > 0$; hence it is eternal expansion. Note that this is the case even if $K < 0$ as long as $K > U_s$.

Case (3b) $K < U_s$: Here, two cases are possible. Let the two solutions for $U(x) = K$ be x_1, x_2 where $x_1 > x_2$. If the point mass, that is, the Universe, is on the left side of the potential ($x \leq x_2$), the motion is expansion and contraction. If it is on the right side of the potential ($x \geq x_1$), it is contraction, rebounce, and expansion. This is the case in which no Big Bang occurred because x never reaches 0.

Case (3c) $K = U_s$: Here, three cases are possible. Let $x_s = R_s/R_0$ be the stationary point where $U(x_s) = U_s$.

(3ca) $x < x_s$: The universe is expanding but the expansion rate is slowing down to vanish. It takes infinite time to reach $x = x_s$.

(3cb) $x > x_s$: It is expansion. But at $t = -\infty$, the point started from $x = x_s$.

(3cc) $x = x_s$: This is a stationary point. The universe does not expand nor contract. This is Einstein's stationary solution. The position is unstable but since it takes infinite time to reach or to leave from the stationary point, observationally it is possible to have a (meta) stable universe.

Observations show that $K = 0$, $\dot{R} > 0$, $\ddot{R} > 0$, $\Omega_{\text{rad}} \approx 0$, $\Omega_m \simeq 0.3$, and $\Omega_\Lambda \simeq 0.7$. This means that the case (3a) applies to our Universe and we are in the acceleration stage, that is, $x(t_0) = 1 > x_s$. By setting $\Omega_{\text{rad}} = 0$, we can calculate the stationary point to be $x_s = (\Omega_m/2\Omega_\Lambda)^{1/3}$. To convert x_s to time, we use $x \propto t^{2/3}$ in matter-dominant universe. We obtain

$$\begin{aligned} 1+z &= \frac{R(t_0)}{R} = \left(\frac{t_0}{t}\right)^{\frac{2}{3}}, \qquad t_0 = 13.8 \text{ billion years}, \\ \rightarrow \quad t_s &= t_0\sqrt{\frac{\Omega_m}{2\Omega_\Lambda}} \approx 5.9 \text{ billion years or } z_s = 0.76. \end{aligned} \tag{G.5}$$

The time when the universe turned to Λ dominance $(\Omega_\Lambda > \Omega_m)$ can be calculated as

$$t_\Lambda = t_0\sqrt{\frac{\Omega_m}{\Omega_\Lambda}} \approx 8.4 \text{ billion years or } z_\Lambda = 0.39. \tag{G.6}$$

That is, our universe has turned to Λ dominance only very recently.

For the special case of $\Omega_m = 0$, Eq. (G.3) can be solved analytically, which reproduces the approximate shape of the cosmic evolution (see Figure G.1(a)). We

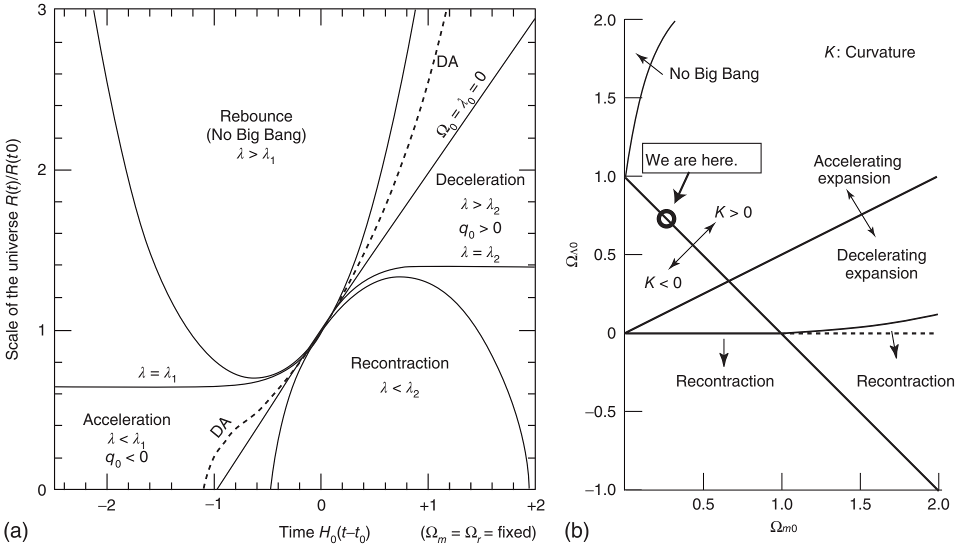


Figure G.2 (a) Evolution routes of the cosmic expansion. The cosmic evolution curves for given Ω_m, Ω_{rad}, and various choice of Ω_Λ are plotted as a function of time in units of $H_0(t - t_0)$. Observations show that our Universe's expansion turned from decelerating to accelerating expansion very recently, which is shown as the line denoted as DA. $\lambda = \Omega_\Lambda/\Omega_m$ and λ_1, λ_2 are solutions of $\max[U] = K = 1 - \Omega_m - \Omega_\Lambda$. λ_2 becomes negative for $\Omega_m, \Omega_{\text{rad}} < 1$. (b) Cosmic global behavior on the Ω_{m0} —$\Omega_{\Lambda 0}$ parameter plane.

show in Figure G.2(b) how the fate of our Universe might have been different by a different setting of the cosmological constant. The dotted line denoted as DA in Figure G.2(a) shows qualitatively the past and predicted future of our Universe by using the observed numbers for the parameters.

Color Plates

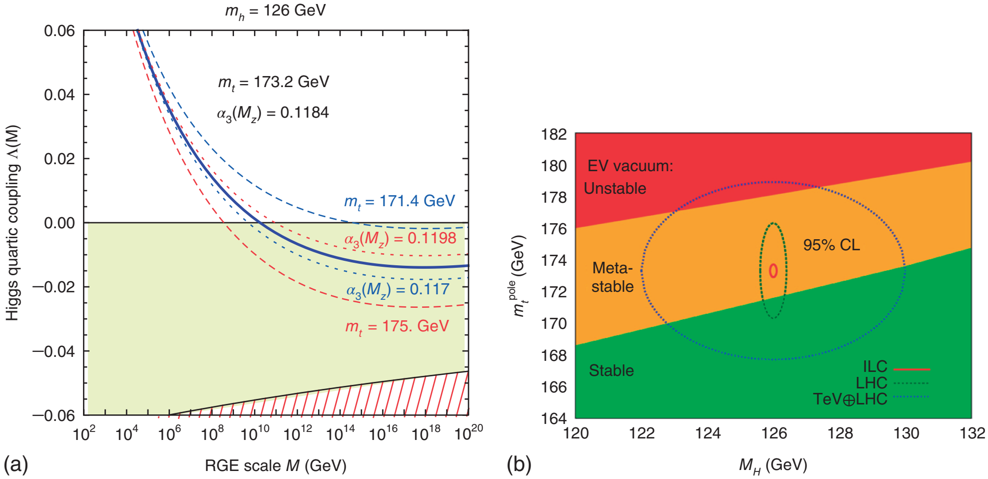


Figure 1.6 (a, b) Renormalization group evolution of the Higgs self-coupling for $m_h = 126$ GeV for the central value of m_t and α_s, as well as for $\pm 2\sigma$ variations of m_t (dashed lines) and α_s (dotted lines). For negative values of λ, the lifetime of the SM vacuum due to quantum tunneling at zero temperature is longer than the age of the Universe as long as λ remains above the region shaded in red, which takes into account the finite corrections to the effective bounce action renormalized at the same scale as λ. (Reproduced with permission of [8, 22].). The 2σ ellipses in the $m_H - m_{\text{top, pole}}$ plane are drawn for the three cases that one obtains from the current top quark and Higgs mass measurements at the Tevatron and LHC, and which can be expected in future measurements at the LHC and at the ILC. The colored area denotes where the SM vacuum is absolutely stable, metastable, and unstable up to the Planck scale. (Reproduced with permission of [16, 23].) (This figure also appears on page 16.)

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

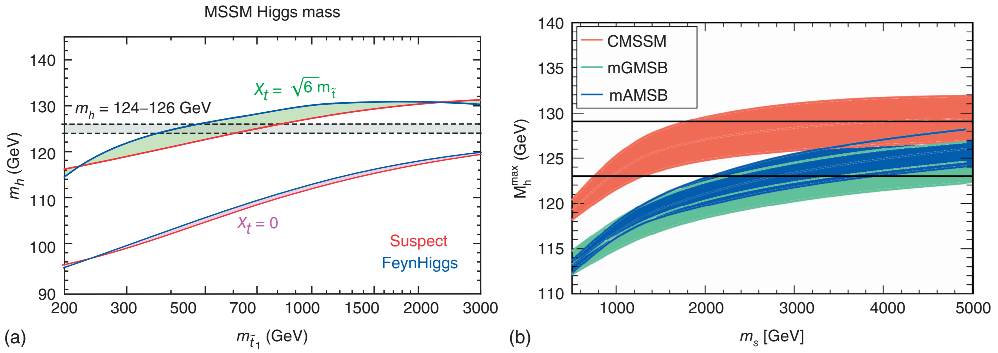


Figure 1.10 (a) Higgs mass as a function of the scalar top mass. Other SUSY parameters were fixed at values guided by naturalness conditions. Large mixing and stop mass are required. (Reproduced with permission of [37]). (b) The Higgs mass shown as a function of $M_S = (m_{\tilde{t}_1} m_{\tilde{t}_2})^{1/2}$ for the various constrained MSSM models. Note that, for reasonable values of $M_S \lesssim 3000$ GeV, simple versions of GMSB, AMSB are ruled out. CMSSM (constrained minimum supersymmetric extension of the Standard Model discussed in Section 5.3) survives. (Reproduced with permission of [38, 39].) (This figure also appears on page 27.)

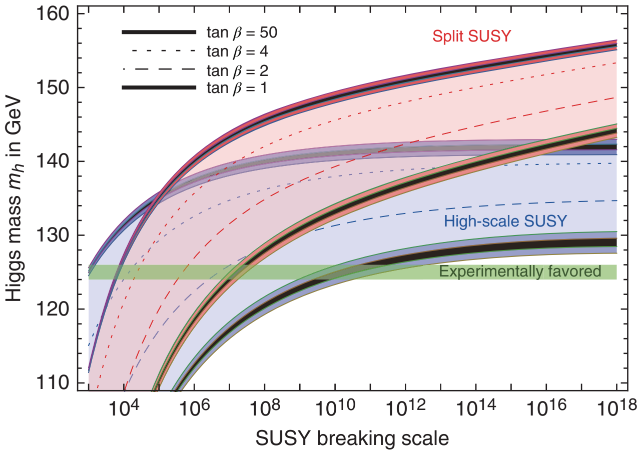


Figure 1.11 NNLO prediction for the Higgs mass m_h in high-scale supersymmetry (blue, lower) and split supersymmetry (red, upper) for $\tan\beta = \{1, 2, 4, 50\}$. The thickness of the lower boundary at $\tan\beta = 1$ and of the upper boundary at $\tan\beta = 50$ shows the uncertainty due to the present 1σ error on α_s (black band) and on the top mass (larger colored band). (Reproduced with permission of [16].) (This figure also appears on page 28.)

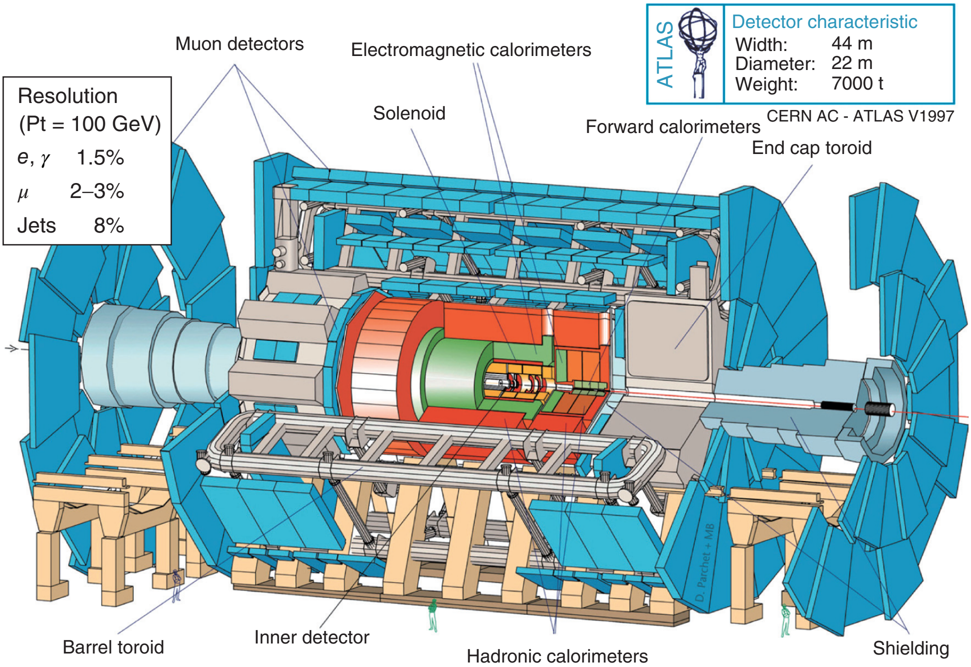

Resolution
(Pt = 100 GeV)
e, γ 1.5%
μ 2–3%
Jets 8%
Muon detectors
Electromagnetic calorimeters
Solenoid
Forward calorimeters
ATLAS
Detector characteristic
Width: 44 m
Diameter: 22 m
Weight: 7000 t
CERN AC - ATLAS V1997
End cap toroid
Barrel toroid
Inner detector
Hadronic calorimeters
Shielding
D. Parchet + MB

Figure 1.20 ATLAS (A Toroidal LHC ApparatuS) is a general-purpose detector with balanced performance including hadronic events. Its magnet is an air-cored one with inner (outer) radius 5 (10) m and length 26 m, and produces a field strength of 0.8 T (2 T at the center). The electromagnetic calorimeter is a liquid argon detector of accordion shape [see Figure 12.35 of [1]]. The overall size is $20\,\phi \times 44$ m, and weighs 6000 tons. Resolutions are good but not particularly good for all measurements. (Reproduced with permission of [83].) (This figure also appears on page 48.)

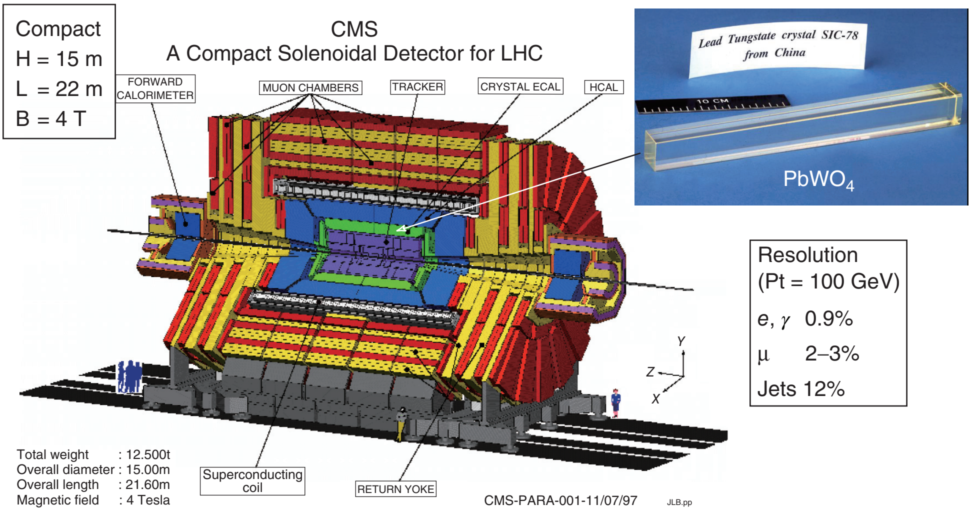

Compact
H = 15 m
L = 22 m
B = 4 T
CMS
A Compact Solenoidal Detector for LHC
FORWARD CALORIMETER
MUON CHAMBERS
TRACKER
CRYSTAL ECAL
HCAL
Lead Tungstate crystal SIC-78 from China
10 CM
PbWO4
Resolution
(Pt = 100 GeV)
e, γ 0.9%
μ 2–3%
Jets 12%
Y
Z
X
Total weight : 12.500t
Overall diameter : 15.00m
Overall length : 21.60m
Magnetic field : 4 Tesla
Superconducting coil
RETURN YOKE
CMS-PARA-001-11/07/97
JLB.pp

Figure 1.21 CMS (compact muon solenoid) is a detector at LHC. The length is about one-half of that of the ATLAS detector but its weight (= 12.5 kton) is twice that of ATLAS. It uses PbWO$_4$ (lead glass) for the electromagnetic calorimeter and realizes a resolution of 0.9% (1.5%) at $P_T = 100$ GeV. The magnetic field strength is 4 T as compared to 2 T of ATLAS. It excels in $e/\mu/\gamma$ physics and is general purpose, but is less versatile for hadron detection. (Reproduced with permission of [84].) (This figure also appears on page 49.)

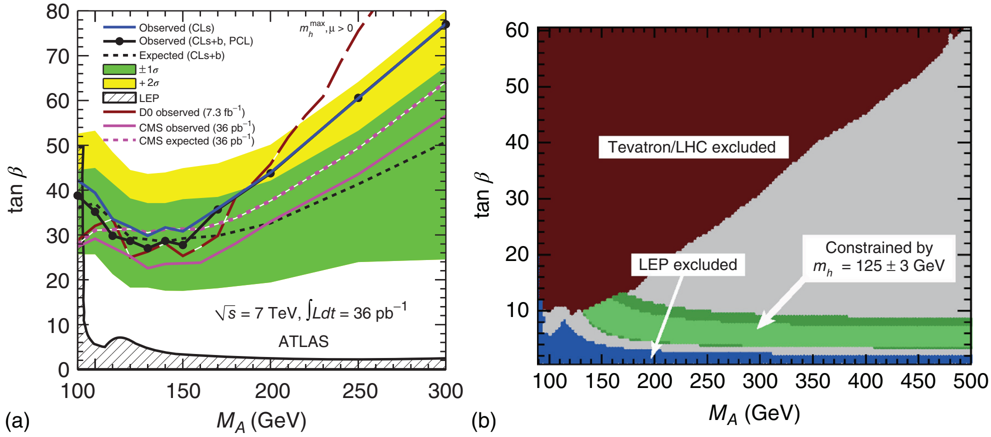


Figure 1.28 (a) Expected and observed exclusion limits in the $m_A - \tan\beta$ plane of the MSSM without the Higgs at 125 GeV. The region above the drawn limit curve is excluded at the 95% CL for the $m_h = \max$ scenario. The dark gray (green) and light gray (yellow) bands correspond to the $\pm 1\sigma$ and $+2\sigma$ error bands, respectively. The exclusion limits from LEP, D0, and CMS are also shown. (Reproduced with permission of [109, 110].) (b) Discovery of the Higgs particle at $m_h = 125$ GeV, if interpreted as the light Higgs of the MSSM constrains the allowed region on the $m_A - \tan\beta$ plane. The green horizontal band shows the region where m_h is compatible with the observed value within errors (125 ± 3 GeV). The brighter green is for the top mass value $m_t = 173.2$ GeV and the dark green corresponds to $\pm 1\sigma = 0.9$ GeV variation. The parameters governing the higher order corrections are taken such that a maximum value for m_h is obtained ($m_h^{\max}$ benchmark scenario). Dark (blue) areas are the excluded regions by LEP and Tevatron/LHC (see (a)). The gray area is the allowed parameter space prior to the latest LHC results. Note: Constraints on the upper edge of the green band is loose. This is still preliminary. If the $m_h = \max$ scenario is modified, the allowed region can be expanded to above significantly. (Reproduced with permission of [111–113].) (This figure also appears on page 57.)

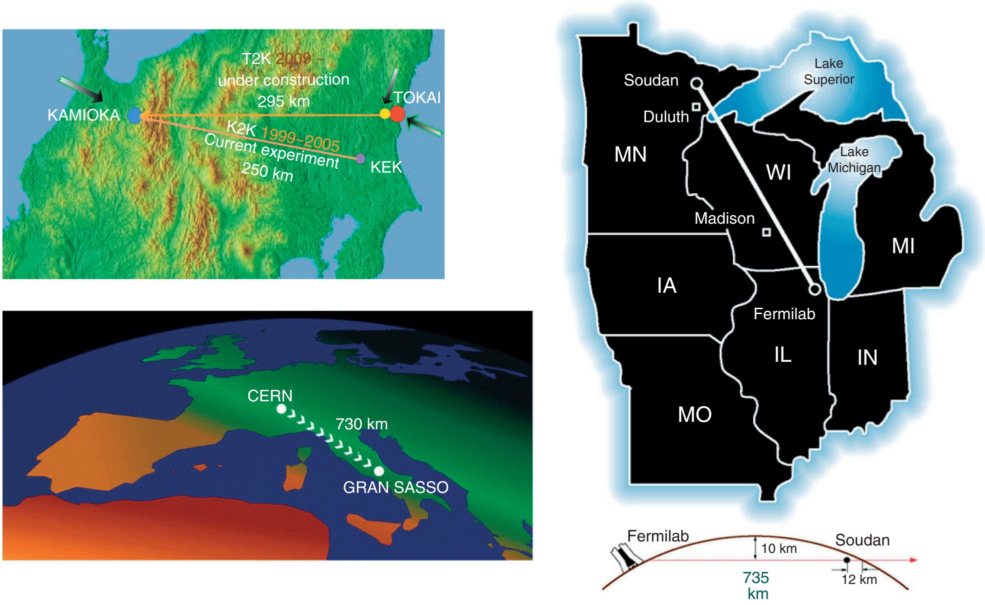

T2K 2009 ~
under construction
295 km
KAMIOKA
TOKAI
K2K 1999~2005
Current experiment
250 km
KEK
CERN
730 km
GRAN SASSO
Soudan
Lake Superior
Duluth
MN
WI
Lake Michigan
Madison
MI
IA
Fermilab
IL
IN
MO
Fermilab
10 km
Soudan
735 km
12 km

Figure 2.10 World neutrino oscillation experiment sites. From top clockwise: K2K (KEK to Kamioka, 250 km), T2K (Tokai to Kamioka, 295 km) in Japan, MINOS (Main Injector Neutrino Oscillation Search, Fermilab to Soudan, 735 km) in the United States, and CNGS (CERN to Gran Sasso in Italy, 730 km) in western Europe. (This figure also appears on page 82.)

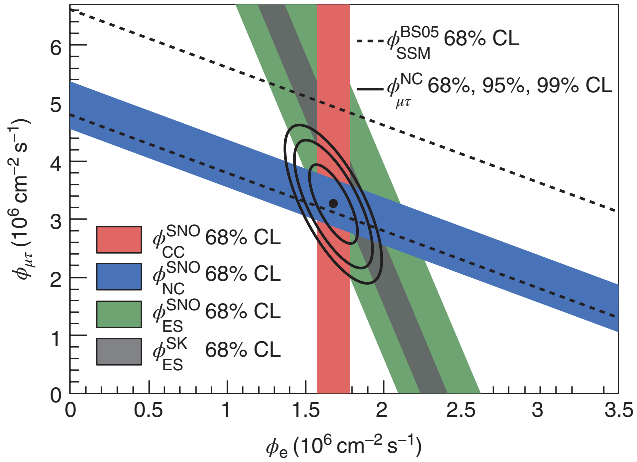


Figure 2.27 SNO data consistency check. The solar neutrino fluxes are measured by three different reactions CC: $\nu_e + D \rightarrow e^- + p + p$, NC: $\nu_e + D \rightarrow \nu_e + p + n$, and ES: $\nu_e + e^- \rightarrow \nu_e + e^-$. Three bands are the SM constraints that express $\phi_{\mu\tau}$ (sum of ν_μ and ν_τ fluxes) as functions of ϕ_e (ν_e flux). The band between the two dotted lines is the prediction of the SSM. The narrow black band of ϕ_{ES} indicates the νe scattering data obtained by SuperK. (Reproduced with permission of [7].) (This figure also appears on page 100.)

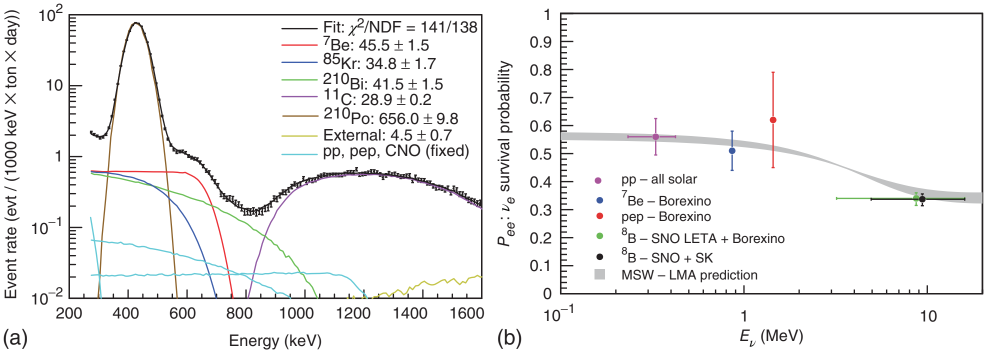


Figure 2.33 Borexino results on ^{7}Be solar neutrino observation. (a) Event rate as a function of the scattered electron energy in $\nu_e + e^- \rightarrow \nu_e + e^-$ reaction in the energy region 270–1600 keV. The solid line shows a Monte Carlo-based fitted spectrum (counts day^{-1} 100 ton^{-1})]. Some, but not all, of the α events have been removed using a cut. (b) Electron neutrino survival probability as a function of energy. The red line and ^{7}Be measurements of Pee are obtained in this experiment. The pp and ^{8}B measurements are also shown. The MSW–LMA prediction is shown for comparison; the band defines the 1σ range of the mixing parameter estimate in [7]. (Reproduced with permission of [231, 232].) (This figure also appears on page 110.)

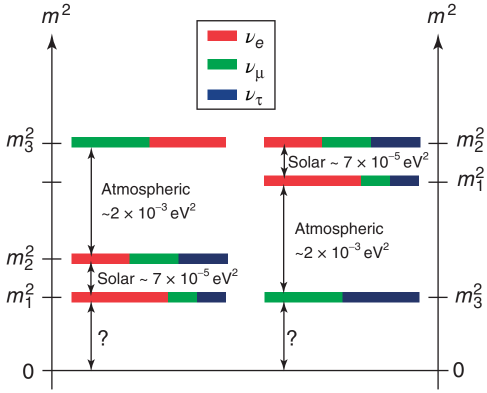


Figure 2.41 Level structure of the neutrino. The left column shows the normal hierarchy and the right one shows the inverted hierarchy. There is also a possibility of degenerate masses ($m_1 \simeq m_2 \simeq m_3 >> \sqrt{|\Delta m^2|}$), which will be described in Figure 2.45. The bands showing levels of the mass are color-coded by mixing ratios of the flavor states obtained from Eq. (2.125). (This figure also appears on page 121.)

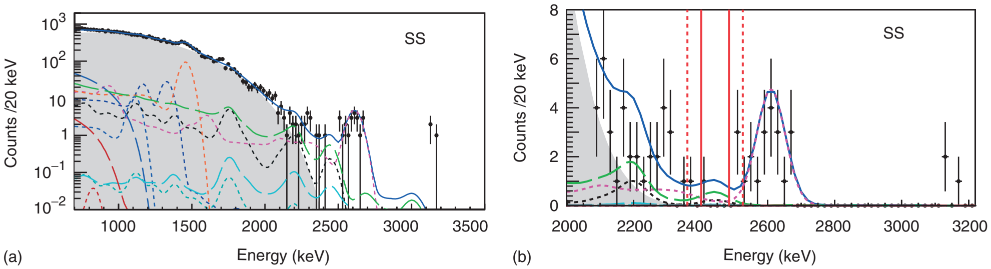


Figure 2.48 (a) Energy spectrum of single-site (SS) events obtained by EXO-200. The best-fit line (solid blue) is shown. The background components are $2\nu\beta\beta$ (gray region), ^{40}K (dotted orange), ^{60}Co (dotted dark blue), ^{222}Rn in the cryostat–lead airgap (long-dashed green), ^{238}U in the TPC vessel (dotted black), ^{232}Th in the TPC vessel (dotted magenta), ^{214}Bi on the cathode (long-dashed cyan), ^{222}Rn outside of the field cage (dotted dark cyan), ^{222}Rn in active xenon (long-dashed brown), ^{135}Xe (long-dashed blue), and ^{54}Mn (dotted brown). (b) Energy spectra in the ^{136}Xe $Q_{\beta\beta}$ region for SS events. The $1(2)\sigma$ regions around $Q_{\beta\beta}$ are shown by solid (dashed) vertical lines. The peak at 2615 keV is due to ^{228}Th. The fit results have the same meaning as in (a). The $0\nu\beta\beta$ probability density function (PDF) from the fit is not visible. (Reproduced with permission of [301].) (This figure also appears on page 136.)

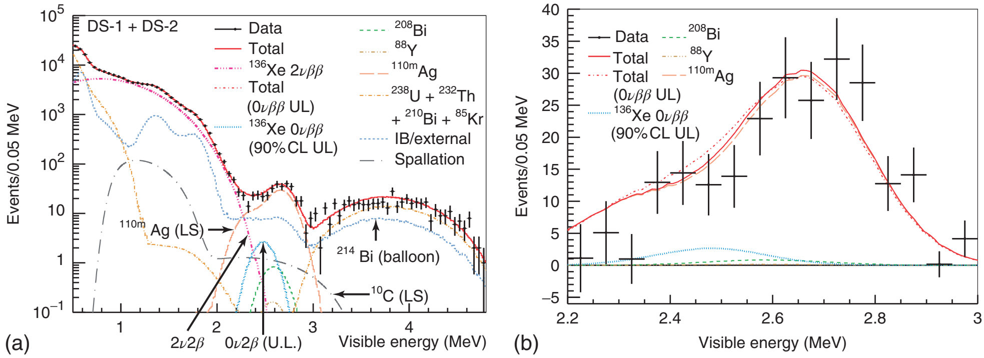


Figure 2.49 (a) Energy spectrum of selected candidate events by KamLAND-Zen together with the best-fit backgrounds and $2\nu\beta\beta$ decays, and the 90% CL upper limit for $0\nu\beta\beta$ decays. The fit range is $0.5 < E < 4.8$ MeV. (b) Closeup of (a) for $2.2 < E < 3.0$ MeV after subtracting known background contributions. (LS) means backgrounds from liquid scintillator and (balloon) from balloon. (Reproduced with permission of [322].) (This figure also appears on page 137.)

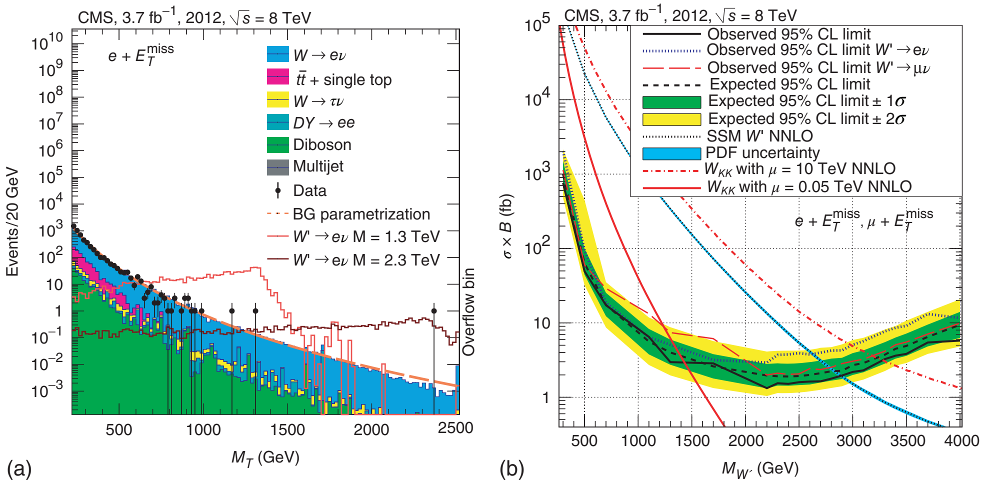


Figure 3.7 (a) Observed transverse mass distributions for the electron channels. Simulated signal distributions for a W' are also shown, including detector resolution effects. The simulated background labeled as "diboson" includes WW, ZZ, and WZ contributions. The top background prediction includes single top and top pair production. The total background prediction from a fit to the simulated transverse mass spectrum in each channel is shown by the dashed line. (b) Limits on the cross section times the single channel-branching fraction $(\sigma \cdot B)$ for heavy W' bosons for the electron and the muon channels. For the individual channels, only the observed limits are shown. For the combination, the observed limit, the expected limit, the 1σ, and the 2σ bands are displayed. All limits are displayed for the single channel-branching fraction. The W' mass limits are for the models of a SSM W' and $W_{KK}^{n=2}$ in split UED (universal extra dimension model) (see Figure 6.2 for interpretation of extra dimensional models). (Reproduced with permission of [386–388].) (This figure also appears on page 181.)

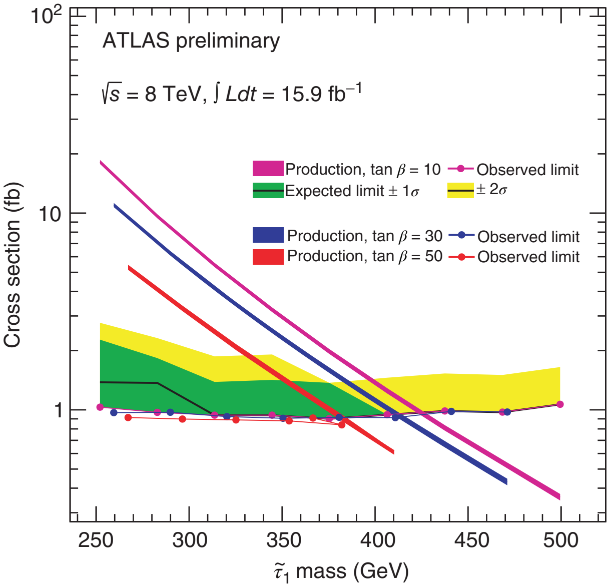


Figure 5.10 Cross section limits as a function of the $\tilde{\tau}_1$ mass in GMSB models. Observed limits are given as solid lines with markers. The theoretical prediction for the cross section is shown as a colored 1σ band. Different colors represent models with different tan β. Expected limits for tan β = 10 are drawn as magenta (top) line with ±1 and ±2σ uncertainty bands drawn in green and yellow, respectively. (Reproduced with permission of [480].) (This figure also appears on page 253.)

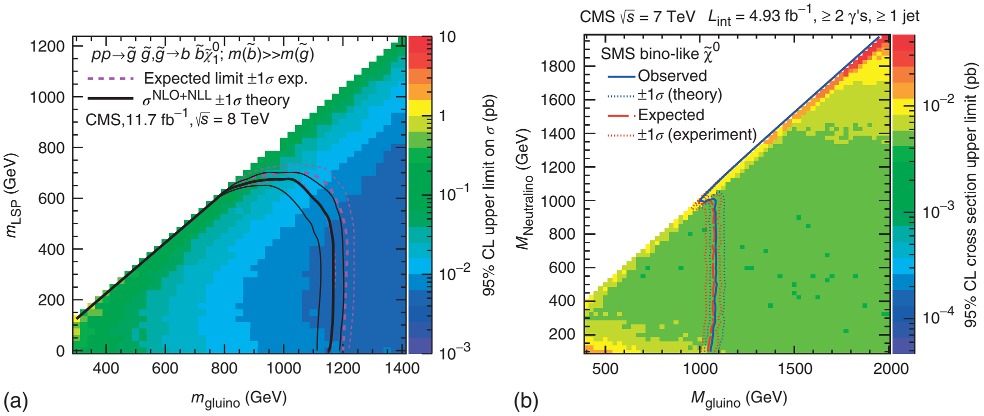


Figure 5.13 Results for simplified models in the form of 95% CL upper limits on the cross section plus overlaid exclusion contours for (a) the bottom [485] analysis as a function of the parent and LSP sparticle masses and (b) the diphoton. In (a), the blue (black) solid thick line indicates the observed exclusion assuming NLO+NLL SUSY production cross section. The blue dotted (black solid) thin lines represent the observed exclusions when varying the cross section by its theoretical uncertainty. The purple dashed thick (thin) line indicates the median (±1σ) expected exclusion. (Reproduced with permission of [485, 486].) (This figure also appears on page 257.)

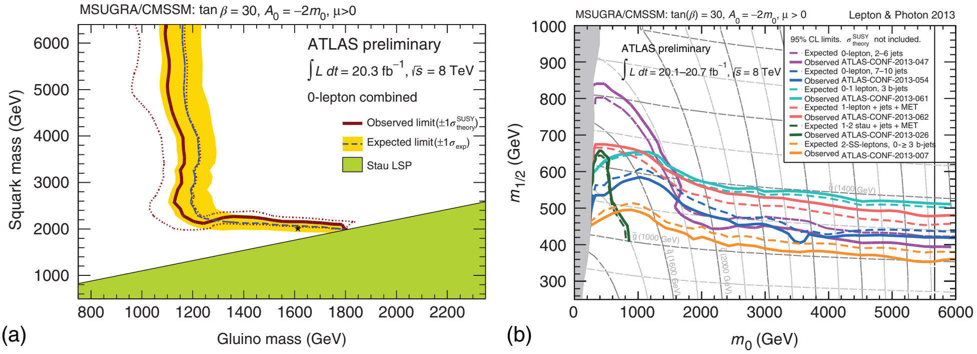


Figure 5.16 Exclusion limits for MSUGRA/CMSSM models with $\tan\beta = 30$, $A_0 = -2m_0$, and $\mu > 0$ presented (a) in the $m_{\tilde{g}} - m_{\tilde{q}}$ plane and (b) in the $m_0 - m_{1/2}$ plane. In (a), the blue dashed lines show the expected limits at 95% CL, with the light (yellow) bands indicating the 1σ excursions due to experimental and background-theory uncertainties. In (b), the observed limits are indicated by medium (maroon) curves, where the solid contour represents the nominal limit, and the dotted lines are obtained by varying the signal cross section by the theoretical scale and PDF uncertainties. (Reproduced with permission of [496, 497].) (This figure also appears on page 260.)

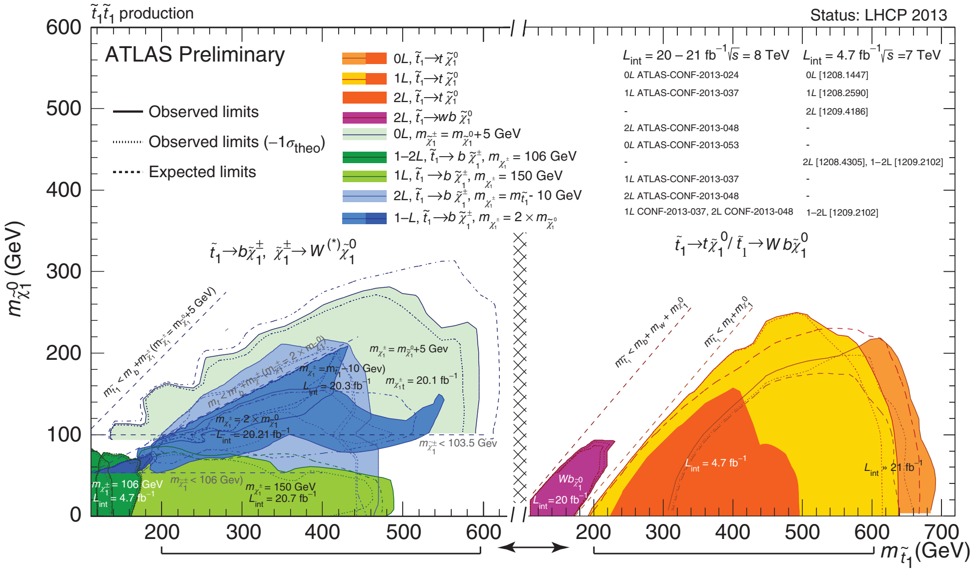


Figure 5.17 Searches for the pair-produced stop with both stops subsequently decaying to top and $\tilde{\chi}^0_1$ excluded the regions in red color. Also shown in blue color are the regions excluded by looking into 1 or 2 leptons + b-jets. Reach for $b\tilde{\chi}^+_1$ analysis is very dependent on $\tilde{\chi}^+_1$ mass but some sensitivity to regions not probed by $t\tilde{\chi}^0_1$ (avoid top in decay chain). (Reproduced with permission of [501, 502].) (This figure also appears on page 261.)

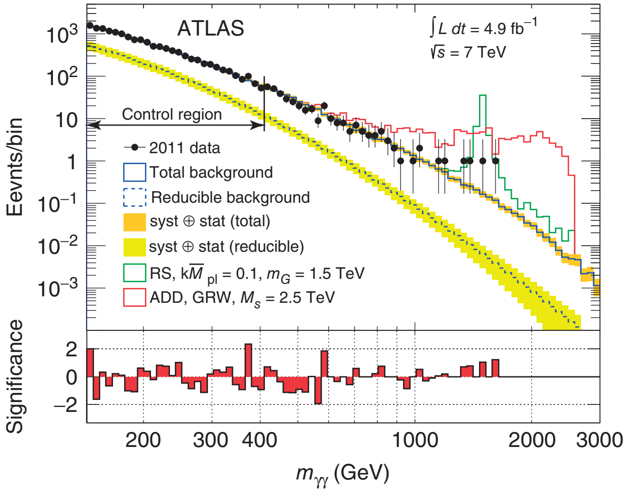


Figure 6.19 Observed invariant mass distribution of diphoton events. Superimposed are the SM background expectation and the expected signals for two examples of RS and ADD models (Reproduced with permission of [596, 597].) (This figure also appears on page 322.)

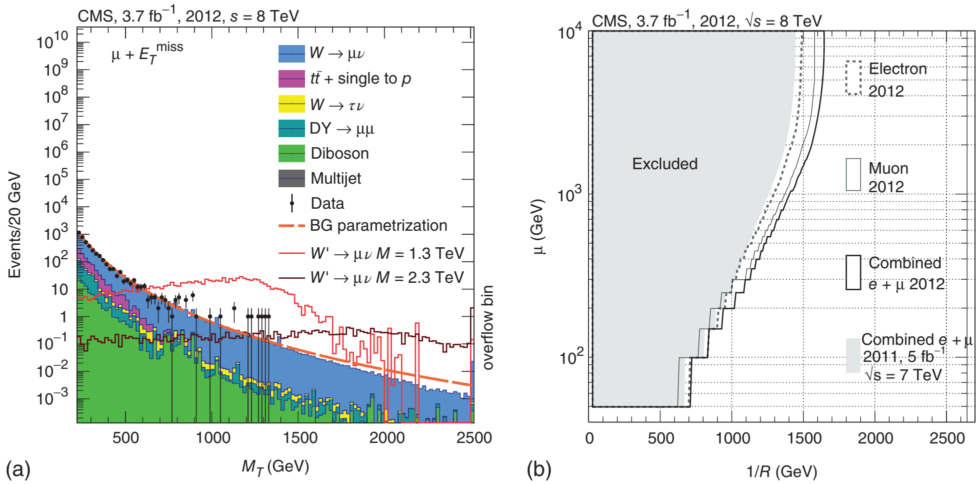


Figure 6.21 (a) Observed lepton-E_T^{miss} transverse mass distributions in the muon channel. The dashed lines show the parametrization of the background. Simulated signal distributions for an SSM W' (which includes $W_{\text{KK}}^{n=2}$) are also shown. The simulated background labeled as "diboson" includes WW, ZZ, and WZ contributions. (b) The 95% CLs on the split-UED fermion bulk mass parameters M (denoted as μ in the figure) and $1/R$ derived from the W' mass limits taking into account the corresponding width of the $W_{\text{KK}}^{n=2}$. For the 3.7 fb^{-1} of 2012 data, the individual limits for the electron and muon channels are shown together with their combination, improving the excluded parameter space based on the 2011 data shown in grey (Reproduced with permission of [601].) (This figure also appears on page 324.)

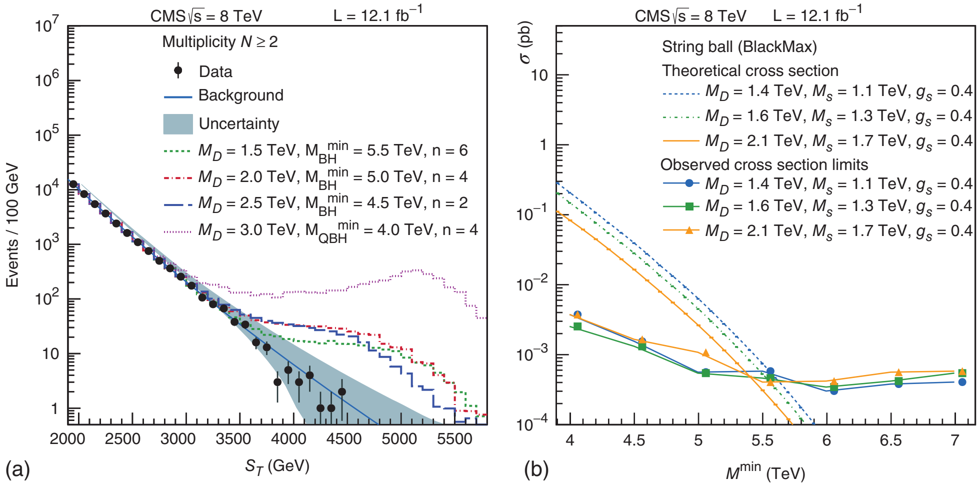


Figure 6.22 (a) Distribution of the total transverse energy S_T with multiplicity $N \geq 2$ objects (photons, electrons, muons, or jets) in the final state. Observed data are depicted as points with error bars; solid line with a shaded band is the background prediction and its systematic uncertainty. Also shown are the expected semiclassical black hole signals for four parameter sets of the BlackMax nonrotating black hole model. Here, $M_{\mathrm{BH}}^{\mathrm{min}}$ is the minimum black hole mass, M_{D} is the multidimensional Planck scale, and n (δ in the text) is the number of extra dimensions. (b) The cross section upper limits at 95% CL from the counting experiments optimized for various black hole parameter sets (solid lines) compared with signal production cross sections from the BlackMax generator (dashed lines) as a function of minimum black hole mass (Reproduced with permission of [610–612].) (This figure also appears on page 327.)

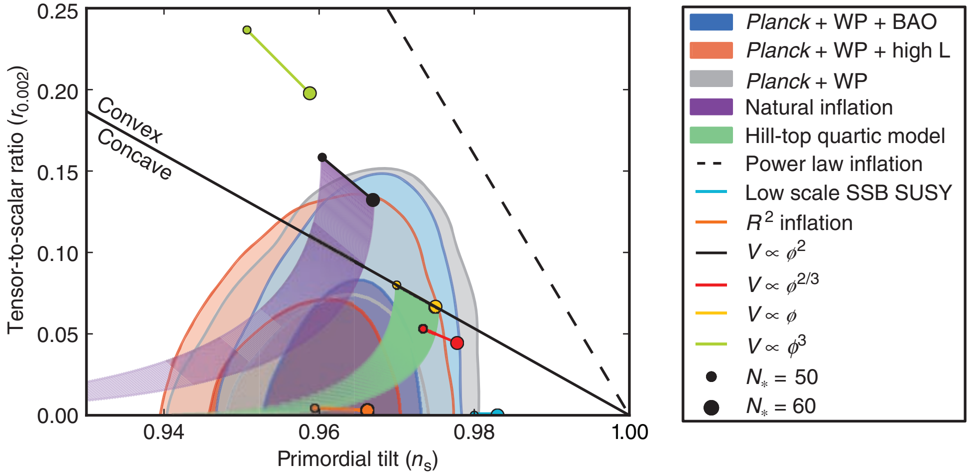


Figure 9.9 Marginalized 68% and 95% confidence levels for n_s and r from Planck + WMAP and BAO data, compared to the theoretical predictions of selected inflationary models. (Reproduced with permission of ESA and [767].) (This figure also appears on page 453.)

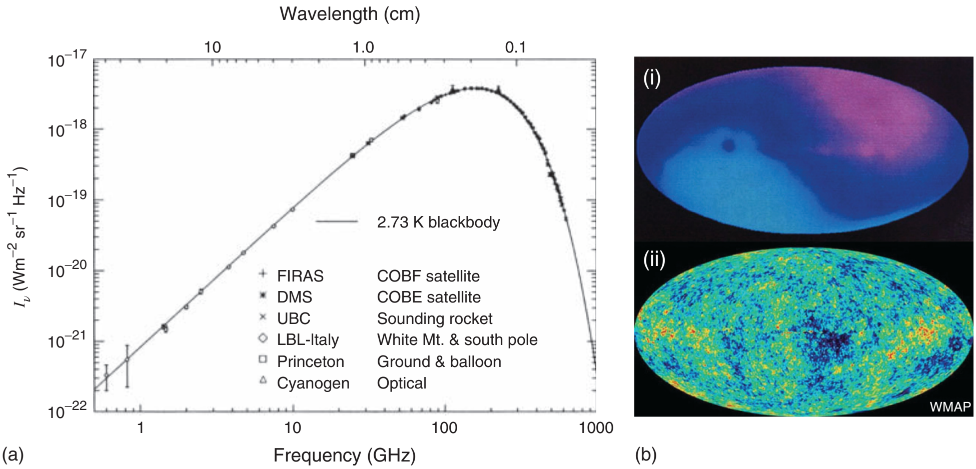


Figure 9.10 (a) CMB spectrum agreeing with the Planck's blackbody radiation formula with error (smaller than the thickness of the line) smaller than 10^{-5}. (Reproduced with permission of [775].) (b) Color-coded anisotropy of CMB. (i): Up to the accuracy of 10^{-3} only a dipole component is observed. [776] (ii): After subtraction of the dipole component, finite anisotropy of $\sim 10^{-5}$ is observed. Figure courtesy of NASA [777]. Figure (a) was obtained by analyzing this anisotropy map. (This figure also appears on page 455.)

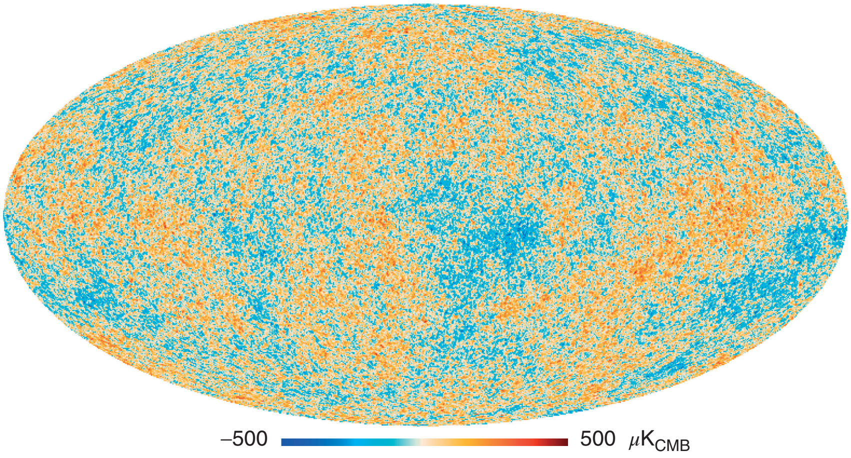


Figure 9.11 Cosmic microwave background (CMB) radiation anisotropy map across the sky obtained by the Planck satellite. The dipole component is subtracted. The intensity of the anisotropy is $O(10^{-5})$. Figure by ESA (Reproduced with permission of [674, 780]). (This figure also appears on page 456.)

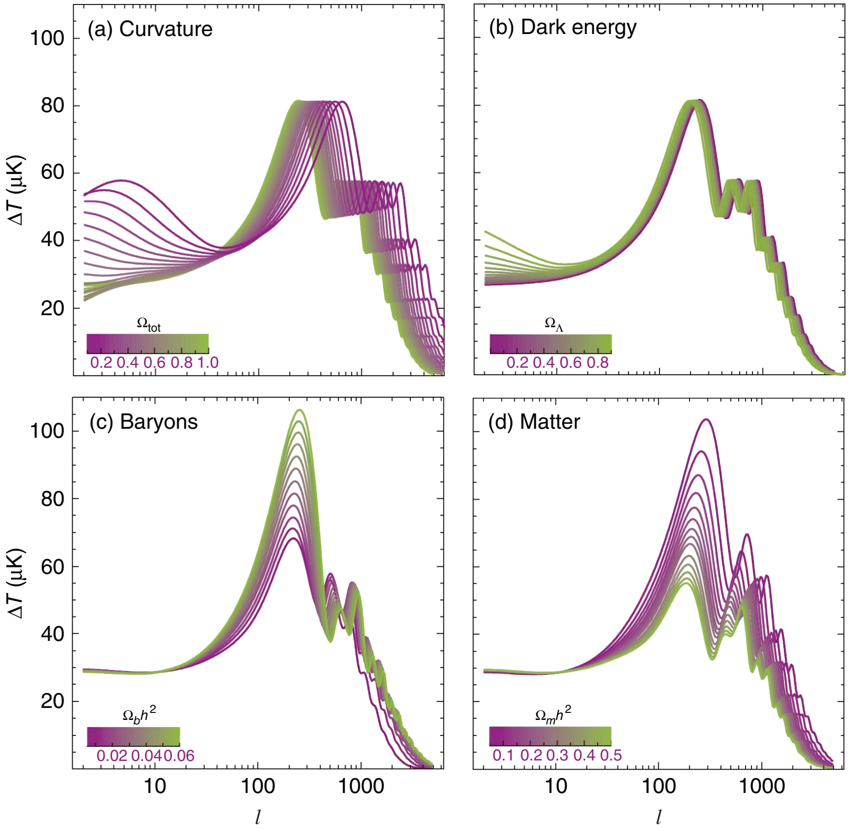


Figure 9.14 Sensitivity of the acoustic temperature spectrum to four fundamental cosmological parameters: (a) The curvature as quantified by Ω_{tot}. (b) The dark energy as quantified by the cosmological constant $\Omega_\Lambda (w = -1)$. (c) The physical baryon density $\Omega_{b0}h^2$. (d) The physical matter density $\Omega_{m0}h^2$, all varied around a fiducial model of $\Omega_{tot} = 1$, $\Omega_\lambda = 0.65$, $\Omega_{b0}h^2 = 0.02$, $\Omega_{m0}h^2 = 0.147$, n = 1, $z_{ri} = 0$, $E_i = 0$. (Reproduced with permission of [791].) (This figure also appears on page 464.)

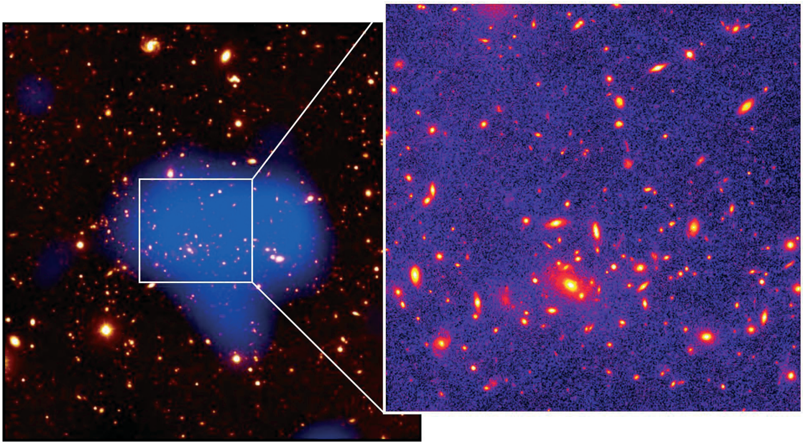


Figure 10.2 Cluster of galaxies MS1054 – 0321 at ~2.5 Gpc distance ($z = 0.831$). Its diameter is ~3 Mpc, and it contains thousands of galaxies. Left: X-ray sources (blue colored) are overlayed with the cluster pictures in the visible range. Right: Enlarged image by the Hubble telescope. (Reproduced with permission of [820–822].) (This figure also appears on page 478.)

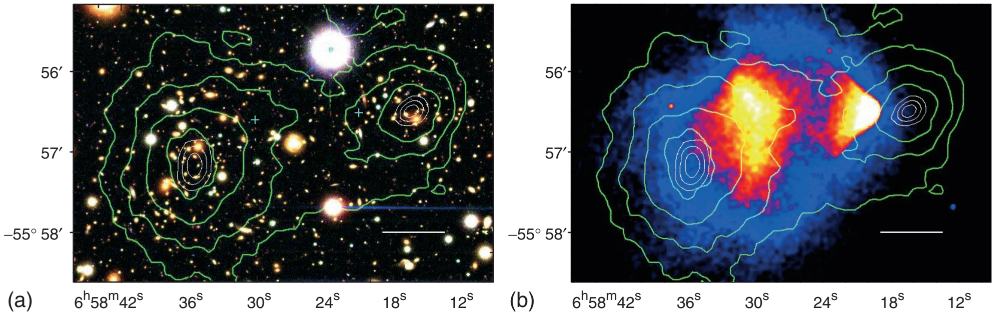


Figure 10.5 (a) Color image of the merging "bullet" cluster 1E 0657-558 obtained with the 6.5-m Magellan telescope in the Las Campanas Observatory, with the white bar indicating 200 kpc at the distance of the cluster. (b) Chandra X-ray image of the cluster. Shown in green contours in both panels are the weak-lensing κ reconstructions, with the outer contour levels at $\kappa = 0.16$ and increasing in steps of 0.07. The white contours show the errors at 68.3%, 95.5%, and 99.7% confidence levels. The two figures are interpreted as follows. The two clusters passed through each other 150 million years ago. The dark matter clump, revealed by the weak-lensing map, is coincident with the collisionless galaxies and lies ahead of the collisional gas. The gas of the smaller "bullet" cluster (right), creating a "bow-shaped shock wave located near the right side of the cluster, lags behind the cluster galaxies." (Reproduced with permission of [828].) (This figure also appears on page 485.)

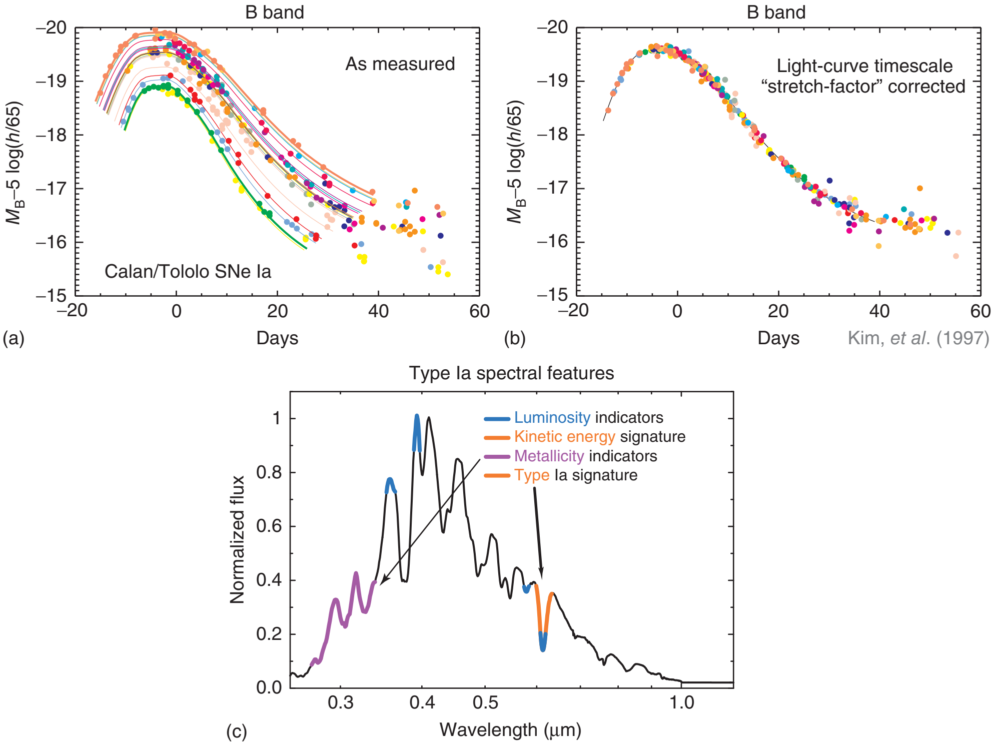


Figure 11.1 (a,b) Light curves as a function of time. The raw spectra show variations but after calibration one obtains a unique absolute luminosity curve which can serve as a standard candle. (c) Typical spectrum of the SN Ia characterized by the absence of the hydrogen line and existence of the Si line. Luminosity and metallicity indicators are also designated (Reproduced with permission of [890, 891].) (This figure also appears on page 514.)

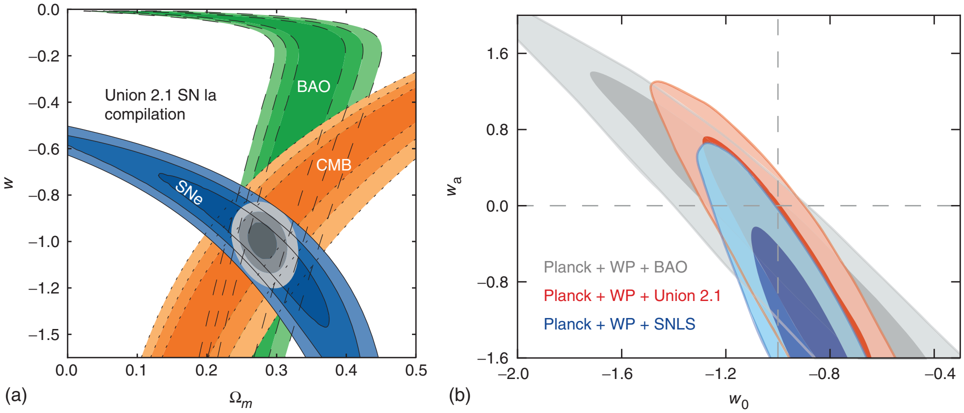


Figure 11.14 (a) 68.3%, 95.4%, and 99.7% confidence regions of the (Ω_m, w) plane. (Reproduced with permission of [905, 906].) (b) 2-D marginalized posterior distributions for w_0 and w_a, for the data combinations Planck+WP+BAO (gray), Planck+WP+Union2.1 (red) and Planck+WP+SNLS (blue). The contours are 68% and 95%, and dashed gray lines show the cosmological constant solution (Reproduced with permission of [123].) (This figure also appears on page 538.)

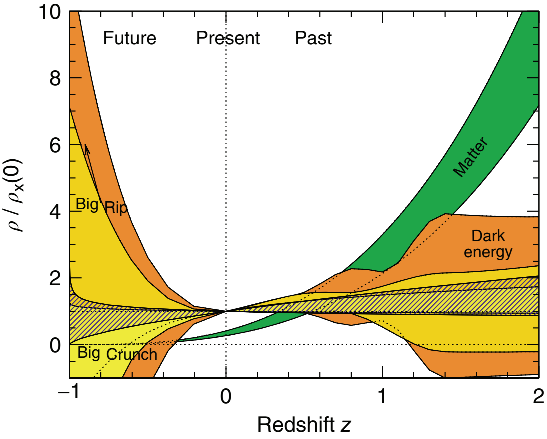


Figure 11.15 1σ constraints on the density of matter and dark energy from SN Ia, CMB, and LSS data, all in units of the current dark energy density (Reproduced with permission of [941].) From inside out, the four nested dark energy constraints are for models making increasingly strong assumptions, corresponding, respectively, to the four-parameter spline, the three-parameter spline, the two-parameter f_∞, w_i case, and the one-parameter constant w case (hatched area). The Universe starts accelerating when the total density slope $d\ln\rho/d\ln(1+z) > -2$, which roughly corresponds to when the dark energy begins to dominate, that is, to where the matter and dark energy bands cross. In the distant future, the Universe will recollapse if the dark energy density ρ_X goes negative and ends in a "Big Rip" if it keeps growing $(d\ln\rho_X = d\ln(1+z) < 0)$. (This figure also appears on page 539.)

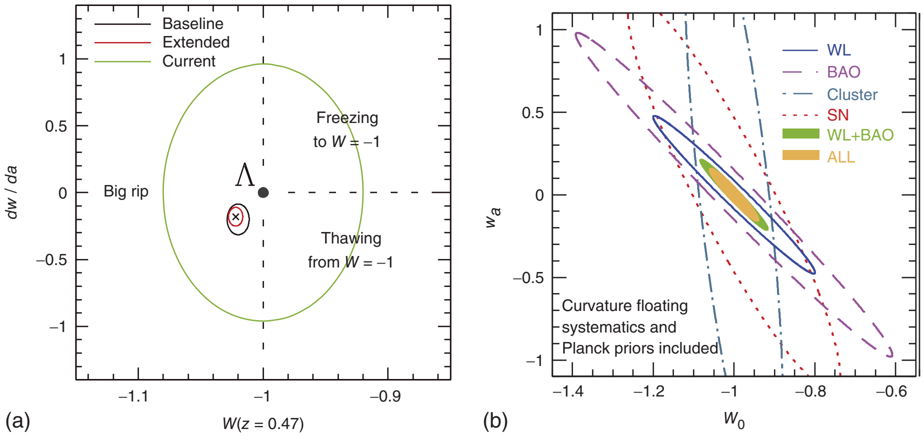


Figure 11.16 (a) Forecast constraints on dark energy parameters from WFIRST-2.4 compared to current knowledge. dw/da is $-w_a$ in the text. Ellipses show $\Delta\chi^2 = 1$ error contours on the value of w at redshift z = 0.47. The green ellipse, centered here on the cosmological constant model (w = −1, dw/da = 0), represents current state-of-the-art constraints from a combination of CMB, SN, BAO, and H0 data, based on Anderson *et al.* (2012). The black ellipse shows the error forecast for the baseline WFIRST-2.4 supernova, galaxy redshift, and weak lensing surveys, combined with CMB data from Planck, a local supernova calibrator sample, and measurements of BAO and redshiftspace distortions from the SDSS-III BOSS survey at $z < 0.7$. The red ellipse shows the "extended" case in which the precision of the WFIRST-2.4 is increased by a factor of two, as a result of a longer observing program. (Reproduced with permission of [940, 944].) (b) Forecasts of constraints on $w_0 - w_a$ region. LSST BAO (baryon acoustic observation, dashed line), cluster counting (dash-dotted line), supernovae (dotted line), WL (weak lensing, solid line), joint BAO and WL (green shaded area), and all combined (yellow shaded area). The BAO and WL results are based on galaxy–galaxy, galaxy–shear, and shear–shear power spectra only. Adding other probes such as strong lensing time delay and higher order galaxy and shear statistics will further improve the constraints (Reproduced with permission of [945, 946].) (This figure also appears on page 540.)

Index

a
Abrikosov's vortex 278, 336
absolute magnitude of star brightness 426
accelerating universe 513
achromaticity 494
acoustic oscillation 459, 461, 462
ADD model, *see also* LED 270
adiabatic 430, 435, 447, 461
– approximation 104
– condition 103, 104
– fluctuation 443, 459, 461
– index 432, 547
adjoint representation 164, 165
Adler-Bell-Jackiew anomaly, *see* chiral anomaly 158
AdS_5: anti-de Sitter space 298
AGB: asymptotic giant branch 364, 518
age crisis 516, 518
Aharonov-Bohm effect 291, 336
Aharonov-Bohm phase 290
ALP: axion-like particle 362, 373
α_T analysis 256, 257
AMSB *see* SUSY models
– AMSB
angular diameter distance 398
angular momentum operator 199
annihilation catastrophe 402
anomalous dimension 246
anomaly, *see* chiral anomaly 350
anomaly mediated symmetry breaking *see* supersymmetry AMSB 232
anthropic principle 522
anti-deSitter space 296
apparent magnitude of star brightness 425–427
apparent luminosity 515
asymptotic freedom 161
ATLAS 2, 31, 48
auxiliary field 215, 217, 219
axial charge 352, 355
axial current 350, 353, 355
axion 329, 356
– ADMX 374–376
– bremsstrahlung 363
– CAST 373
– cold dark matter 363
– coolant 363
– dark matter 329, 363, 376
– decay constant 357
– DFSZ – 361, 363, 376
– hadronic, *see* KSVZ – 361
– helioscope 373
– hot dark matter 363
– invisible – 357, 360, 363
– KSVZ – 361, 363, 373, 376
– microwave cavity detector 374
– misalignment – 494
– PQ
– – charge 358, 360, 362
– – symmetry 356, 357
– – transformation 358
– Primakoff effect 363, 365, 374
– standard – 357
– telescope search 373
– thermal – 373

b
B-L, *see also* Δ (B-L) 171, 172
B-L conservation 229
Baker-Campbell-Hausdorff formula 201
BAO *see also* baryon acoustic oscillation 63, 441, 443
baryogenesis 395, 401
baryon
– –photon fluid 460
– acoustic oscillation 63

Beyond the Standard Model of Elementary Particle Physics, First Edition. Yorikiyo Nagashima.
© 2014 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2014 by Wiley-VCH Verlag GmbH & Co. KGaA.

baryon (*contd.*)
- asymmetry 400, 401, 404, 406–408
- density 388, 401, 464, 465
- number 158, 171
  - non-conservation 348, 349

BB, *see* neutrino beta beam 124
Bessel function 304, 337
β function 160
β function 160, 170, 185, 246, 283, 284
big bang 377
- cosmology 377, 396, 400, 401, 412, 413, 418, 420
- nucleosynthesis 377, 395, 410, 413

Big Crunch 528
big hierarchy 16, 244, 267, 283
- problem 182

big hierarchy problem 18, 283, 285, 299
Big Rip 533
bimaximal 120
black hole 141, 325, 327
- classical – 326
- Hawking radiation 326, 327
- Hawking temperature 325
- production 326, 327
- quantum – 327, 328
- Schwarzschild radius 325

Bolometer 505
Boltzmann equation 394, 406, 429, 447
Bose condensate 5, 420
brane 232, 245
bulk 232, 245
bullet cluster 485

C

Cabibbo-Kobayashi-Maskawa matrix 114, 120
carbon deflagration 514
Casimir
- effect 191
- operator 243

D: charge coupled deviceI 514
CDM, *see* cold dark matter 437
Chandrasekhar mass 140, 548
charged current interaction 81, 85, 86, 99, 101
chemical potential 384, 385, 389, 406, 407, 410, 411, 416
Chern-Simons number 339, 342, 349
chiral
- anomaly 329, 348, 350, 352, 356, 359, 361
- charge 357
- current 355, 358
- fermion in the brane 281
- symmetry 159, 351, 353
- transformation 351–353, 356, 357

Christoffel symbol 560
CKM, *see* Cabibbo-Kobayashi-Maskawa 3
CKM matrix 3, 20, 354
CKM mechanism 231
Clifford algebra 274
cluster counts 537, 539, 540
CMB 63, 377, 379, 387, 389, 395, 413, 414, 417, 422, 454 *see* cosmic microwave background
- anisotropy 455, 456, 458, 460
- B-mode 452, 468, 470, 471, 473
  - measurement 474
  - spectrum 452, 455
- dipole component 455
- E-mode 452, 468, 470, 471
- Planck 450, 453, 456, 457, 519
- power spectrum 452, 454, *see also* CMB anisotropy 456, 457, 461
- quadrupole component 467, 468, 471
- T-E cross-correlation 452, 472
- tensor to scalar ratio 474
- WMAP 311, 450, 455, 519

CMS 2, 48, 49
CMSSM see SUSY models
- CMSSM

coincidence problem 418, 523
cold dark matter 388, 437, 491
collective symmetry breaking, *see* little Higgs model 33
color anomaly 358, 359, 362
color magnitude diagram, *see* Hertzsprung-Russel diagram 364
comoving
- coordinate 379, 380, 384–386, 396–398
- distance 379–381, 396–398, 482
- frame 483
- volume 490

compact group 551
compactification 274, 276, 296, 302
completeness relation 192, 197
complex mass 354
composite
- model 156, 184
- particle 19
- scalar 183

Concordance Model 377, 388, 466, 519, 520
conformal Newtonian gauge, *see* Newtonian gauge 562
connected
- infinitely – 552
- simply –, *see* simply connected 553

connectedness 552
connection 559
constrained MSSM *see* SUSY models CMSSM 231
continuity equation 429, 462
convective equilibrium 545
convergence 482, 484, 488
Cooper pair 28, 184, 336, 351
core collapse 141, 146
coset group 557
cosmic
– budget 475
– destiny 539, 567
– equation 378, 385
– parameter 386–388
– phase transition 378
– variance 457, 465
cosmic destiny 396
cosmic ray background 90
cosmological constant 293, 297–299, 383, 384, 387, 392, 418, 435, 464, 520, 532
– problem 418
cosmological parameter 441, 451, 454, 464, 466, 467, 516
cosmological principle 378, 379, 384, 396, 454
Coulomb barrier 412
coupling strength
– fermion *see* flavor hierarchy in ED 282
– gauge field in 5D 288
– KK graviton in RS model 306
– scalar field in 5D 272
covering group 554
CP
– asymmetry 123
– eigenstate 128
– invariance 20, 24
– phase 121, 128, 129
– violating term 115, 117
– violation 115, 116, 121, 126
CPT theorem 403
critical density 383, 387
curvature 378, 379, 385, 387, 398, 423, 424
– of the universe 464, 520
– parameter 387, 388
– perturbation 427, 459
– spatial– 563
– term 392
custodial $SU(2)$ symmetry 20

d
D-brane 269
dark age 417, 466
dark energy 377, 384, 388, 396, 464, 513
– component 465
– models 532
dark matter 229, 244, 308, 377, 388, 413, 415, 435–437, 440, 448, 454, 458, 475
– candidates 493
– – axion 494
– – MACHO 494
– – neutrino 494
– – WIMPs 494
– component 466
– detection method 495
– large scale structure 488
– searches
– – accelerators 497
– – astrophysical 495
– – underground 496
– – WIMPs, *see* WIMPs wind 505
Daya Bay 119
de Sitter space 297
deceleration parameter 387
decoupling, *see* photon decoupling 435
degeneracy condition 230, 237, 241
degeneracy problem 537
degenerate state 514
degenerate vacua 554
delayed decay mechanism 404
Δ (B-L) 350, 404, 407
δ function 195, 211
Δm^2_{21} 120
Δm^2_{31} 120
Δm^2_{32} 87, 120
Derrick's theorem 339, 341
DGLAP evolution equation 41
diffusion length 436, 462, 463, 465
Dirac
– equation 66
– field 66, 114
– mass 65, 68
– neutrino 63, 66, 67, 71
– particle 65, 68, 72, 128
– phase 114
Dirichlet condition 269
distance modulus 427, 515, 516, 537
Doppler effect 455, 457, 461, 462, 474
double beta decay 126
– 0ν mode 127
– 2ν mode 127
– EXO200 134, 136
– KamLAND-Zen 134, 136, 137
doubly connected 553
Dp-brane, *see* D-brane 269
Drell-Yan process 178

e
e-like event 80, 81, 90
E_6 174, 177, 178, 180
ECC, *see* emulsion cloud chamber 85, 125
ED, *see* extra dimension 270
– searches for ED
– astrophysical 314
– collider experiments 316
– – graviton exchange 321
– – graviton production 316, 318, 319
– cosmic diffuse gamma rays 316
– early universe 315
– neutron star 315
– SN1987A 314
– TeV^{-1} ED
– – ADD model 322, 323
– – RS model 322, 323
– – UED model 324
– split fermion 285
– TeV brane 298, 299, 305
– UED 270, 308, 313
Einstein equation 293, 378
– in ED 293
Einstein ring 481
Einstein tensor 293
Einstein's stationary solution 568
elastic ribbon model 333
electroweak
– mixing 4
– phase transition 348
– precision data 17
– scale 168, 182
ellipticity 484–487
emulsion 85, 86
– cloud chamber 85
entropy 390, 391, 404, 409, 430, 435, 459, 489, 490
equation of state 385, 389, 390, 431, 533, 536, 545
– parameter 524, 529, 530, 532–534, 536
equipartition 529
equivalence principle 193, 559
equivalence theorem 39
equivalent W approximation 40
escape velocity of Milky Way 499
eternal expansion 568
Euclidean action 346
Euler equation 429
event horizon 399
EXO200 *see* double beta decay
– EXO200–200 136
extra dimension 19, 180, 182, 184, 267
– ADD model 270, 313
– anti-domain configuration 282, 283
– anti-kink 279, 280
– brane 269, 279
– – tension 298
– bulk 269, 279, 305
– bulk action 293
– chiral fermion 273
– domain 279, 281, 282
– – configuration 278, 280, 281
– – coupled fermion 279, 285
– – wall 278
– fat brane 279, 285
– gauge field 287
– gauge-Higgs unification 288, 289, 292
– gravitational field 292
– – decomposition of 294
– gravity 292
– KK graviton 549
– KK tower 270, 283
– – gauge field 288
– – graviton in warped ED 302
– – scalar field 272
– LKP 308, 310, 312, 313, 323
– $\overline{M}_5$: Planck mass in 5D 245, 297
– orbifold *see* orbifold
– Planck brane 298, 299, 302
– Planck mass in D dimension, *see* M_D 293
– $R^{(5)}$: Ricci curvature in 5D 297, 301, 307
– radion 296, 299, 300
– Randall-Sundrum model 297, 307, 313
– – RS1 scenario 299
– – RS2 scenario 300
– S^1/Z_2 275
– scalar field 270
– searches, *see* searches for ED 313
– – KK conservation rule 308
– – KK parity 310
– – split – 324
– – with gravity 312
– – without gravity 312
– volcano potential 303
– warp factor 297, 298, 305
– warped – 270, 296
– Z_2 275
– – operation 291
– – parity 275, 279, 282, 310
– – symmetry 275, 276, 281, 285, 291, 292
– Z_2' operator 277

f
family
– family problem 64, 156
– family structure 158, 160
Fayet-Iliopoulos model, *see* D-term breaking 221

FCNC 19, 30, 175, 246
Fermat's principle 479
field strength 209
fifth force 300
fine tuning 17, 18, 30, 182–184, 225, 245, 265
– problem 530
first law of thermodynamics 384, 390
flat direction 223
flatness problem 418, 423
flavor changing neutral current, *see* FCNC 19
flavor eigenstate 73, 75, 76, 120
flavor hierarchy 268, 282, 285, 286, 314
– in ED 282, 283, 285, 286, 318
form factor 19, 71, 184
formula 41
freeze out 489–491
– temperature 490–492
Friedmann equation 386, 387, 389, 393, 394, 396, 518
fundamental representation 162–164, 177

g
galaxy formation theory 436
Gamow 412
Gamow's criteria 394
gauge
– –Higgs unification, *see* extra dimension 289
– freedom in GR 294
– hierarchy, *see* big hierarchy 244
– sector 4, 22
– transformation
– – in GR 427, 562
– unification 185
gauged away 5, 34
Gauss's law 268
Gell-Mann matrix 34
general gauge mediation *see* supersymmetry GGM 231
general relativity 294, 559
generation puzzle *see* family problem 64
geodesic equation 559
geodesic line 381
GIM mechanism 185
Ginzburg-Landau free energy 335
globular cluster 364, 365, 434, 437, 517
– age of 517
gluon 166
gluon fusion 39, 43, 45, 46, 53
– *see* SUSY models
– GMSB 231
Goldstino 222, 223
Goldstone bosons 24
GPS:global positioning system 82
GR, *see* general relativity 294
graded algebra 199
grand unification 14, 18
Grand Unified Theory 18, 155, 161
Grassmann number 191, 194–196, 198–200, 203, 210
– Jacobian 195
gravitational lens 440, 452, 455, 468, 473
gravitational lensing, *see* lensing 476
gravitational wave 427, 467, 470, 471, 474, 562, 563
– primordial – 451, 474
graviton 267, 269, 292
– in extra dimension 292
graviton *see also* extra dimension graviton 267
GRB: gamma ray bursts 537, 539
GUT, *see* Grand Unified Theory 155
GUT scale 18, 182, 183

h
Hamilton's principle 340
harmonic gauge 294
harmonic oscillator 280
Harrison-Zeldovich isocurvature Spectrum 444
HDM, *see* hot dark matter 437
helium flash 367
helium ignition 365
Hertzsprung-Russel diagram 364, 517
HI 477
hidden sector 224, 231, 232, 236, 241, 245, 246
hierarchy 268
hierarchy problem 19, 155, 182, 184, 225, 241, 267, 268, 299, 305
Higgs 1
– charged – 56
– colored – 403
– coupling strength 53
– decay branching ratio 8
– decay width 7, 8, 44, 53
– detection method 48
– discovery of 51
– in MSSM 22, 55
– mass 26, 27, 36
– mass spectrum 24
– quadratic coupling 19, 32
– quartic coupling 10, 15, 17, 19, 20, 32
– sector 1, 2, 4, 6, 16, 22, 25, 40
– spin-parity 52
Higgsstrahlung 39
homomorphic 558

homotopy 551
homotopy class 337–339, 341, 343, 551, 552, 555
horizon 421, 422
horizon problem 418, 421
horizontal branch 364, 373, 518
horizontal symmetry 64
Hosotani mechanism 289, 291
hot dark matter 437, 490
Hubble
– constant 398, 427, 449
– damping 531
– diagram 515, 516
– distance 535
– expansion 530
– parameter 534
– time 393, 421, 519, 526
hypercharge 176–178, 406

i

I,Q,U,V, *see* Stokes parameter 468
IGM: intergalactic medium 401, 417
imaginary time 346
inequivalent 194
inflation 394, 400, 522, 529
inflation model 387, 418, 435, 448, 450, 453, 454, 457, 513
inflaton 418–421, 428, 435, 448–450
– mass 420
– potential 450, 451
infrared cutoff 283
initial condition
– acoustic oscillation 461
– density fluctuation 462
instanton 329, 340, 345, 348–350, 354, 355, 406
Integrated Sachs-Wolf effect, *see* ISW 459
internal background 92
ionization cooling 126
IR brane: infrared brane, *see* TeV brane 299, 302
irreducible representation 205
ISM: interstellar medium 475
isocurvature fluctuation 461
isomorphic 558
isothermal fluctuation 461
ISW 459, 461, 465
– early 459
– late 459

j

Jacobi identity 200
Jacobian 338
Jacobian peak 50, 181
Jarlskog factor 116
Jeans length 432, 433, 436, 448
– neutrino 436, 447
Jeans mass 433, 434
– baryon 434, 437
– neutrino 436

k

Kähler potential 226
Kaluza-Klein theory 267
KamLAND-Zen 137
kink 278, 332, 334
KK: Kaluza-Klein *see* extra dimension KK 270

l

L to B conversion 406
L-R mixing 240, 248, 251
ΛCDM model 388, 417, 436, 442, 447, 448, 451, 452, 454, 457, 466, 519, 533, 537
Λ dominated universe 392, 393, 399, 418
Landau pole 14
Landau-Zener formula 106
Lane-Emden equation 546
large scale structure 436, 437, 443, 448, 459
last scattering 395, 414, 417, 454, 458, 465, 467, 468
latent heat 418
lattice QCD 15
LED:large extra dimension 270
left-right symmetric model 69
left-right symmetry 174, 175, 178, 179
lens equation 480, 481
lensing 476, 479
– effect 494
– strong 481
– transfer matrix 482, 483, 488
– weak 477, 481, 484–486
leptogenesis 121, 404, 407
lepton asymmetry 400
lepton number 158, 171
leptoquarks 166
LHC 2
Lie algebra 191, 198, 199, 553
Lie group 551
lightest KK particle *see* extra dimension LKP 308
lightest supersymmetric particle *see* SUSY models LSP 229
little hierarchy 16, 308
little hierarchy problem 18, 32, 59
little Higgs model 19, 31, 179
– collective symmetry breaking 33
– littlest Higgs model 38
– simplest model 38

longitudinal polarization 8, 13
Low Scale Techni-Color model 30
luminosity distance 398
luminosity function 43
luminosity of WW collision 41
Lyman α 417, 466

m
M/L: mass to luminosity ratio 436, 479
magnetic flux soliton 335
main sequence 62, 364, 517, 518, 545
Majorana
– condition 66, 194, 197
– fermion 228, 238
– field 65, 66, 194, 196
– heavy –neutrino 404, 405, 407
– – mass 407
– Lagrangian 196
– magnetic moment 72
– mass 65, 128, 134, 135
– neutrino 61, 63, 67, 68, 72, 126
– – test of 126
– particle 62, 65, 128
– phase 76, 114
– spinor 192, 199, 200
Majoron 67
manifold 551
mass eigenstate 73, 74, 76, 101–105, 107, 115, 120, 130
mass hierarchy 69, 121, 126, 127, 267
– problem 64
matter density 464
matter dominated 392, 393, 399, 414
matter eigenstate 105
matter oscillation 100
matter-antimatter asymmetry, *see* baryon asymmetry 400
matter-radiation equality 395, 414, 428, 435, 436, 444–446, 448, 513, 523
Maxwell distribution 498
Maxwell-Boltzmann factor 411
M_D: Planck mass in D dimension 232, 268, 293
messenger 242, 243
metastable vacuum 15
metric tensor 427
Michel parameter 69
Mikheyev-Smirnov-Wolfenstein effect, *see* MSW effect 103
minimum SUGRA *see* SUSY models
– mSUGRA 231
minimum supersymmetric extension of the Standard Model, *see* MSSM 20
modulus field 300
mono jet 263, 319, 320, 497
monopole 340
monopole probelm 418, 424
MSSM 20, 25, 223, 224, 226, 236, 237, 239, 241, 249, 283
– brane 245, 246
– particles 241
MSW effect 100, 103
MT2 analysis 257
μ-like event 80, 90

n
Nambu-Goldstone Boson 29, 32, 33, 35, 355–357
Nambu-Goldstone vacuum 351
natural isotope 92
naturalness 27, 263, 265
Neumann condition 269, 302
neutral current interaction 99, 101
neutrino 61
– active – 99
– beta beam 124
– charge distribution 70
– CNGS 82, 85, 86
– decoupling 395, 408, 412
– electromagnetic interaction 70
– heavy – 74
– magnetic moment 71
– mass 62, 409
– – inverted hierarchy 121, 129, 130
– – normal hierarchy 121, 129
– – quasi degenerate 121, 129, 130
– – seesaw mechanism *see* seesaw mechanism 68
– mass hierarchy 120, 121
– mass matrix 65, 66
– mixing 73, 75
– neutrino factory 124, 125
– NUMI 84
– right-handed 62
– sterile – 63, 81
– super-beam 124
– supernova 136
– telescope 97
– temperature 409
neutrino oscillation 61, 76
– accelerator experiments 82
– atmospheric – 79
– Borexino 109, 110
– bronze channel 124
– CP term 122, 123
– evidence 80, 100
– golden channel 101
– IMB 79, 97, 136

neutrino oscillation (*contd.*)
– K2K 82, 86, 87
– Kamiokande 87
– long baseline experiment 76, 123, 126
– matter effect, *see* MSW effect 101
– MINOS 82, 84, 86, 87
– mixing angle 120
– – θ_{13} 122
– – $\sin^2\theta_{12}$ 109
– – $\sin^2\theta_{13}$ 119
– – $\sin^2\theta_{32}$ 87
– – $\sin^2\theta_{atm}$, *see* $\sin^2\theta_{32}$ 87
– – θ_{13} 118
– NOvA 123
– OPERA 85
– oscillation curve 81, 109
– reactor, *see* reactor experiment 111
– Schrödinger equation 101
– short baseline experiment 76, 116, 117, 121
– silver channel 125
– solar, *see* solar neutrino 93
– SuperK 80, 86–88, 92
– T2K 82, 123
– three-flavor – 114
– two-flavor – 76
neutron electric dipole moment 354
Newtonian gauge 427, 446, 562, 563
Newtonian limit 430
NF, *see* neutrino factory 124
NGB, *see* Nambu-Goldstone Boson 32
no big bang 520, 568
Noether current 334, 352, 353
non-linear **σ** model 32
non-SUSY GUTs 174
normalcy temperature 316
nuclear form factor 503
nuclear matrix element 128, 131, 135
nucleosynthesis 526

o
O'Raifeartaigh model, *see* F-term breaking 221
old standard model in cosmology 519
onion structure 145
onion structure of pre-supernova 145
optical depth 466
orbifold 275–278, 281, 291, 292, 298, 302, 303, 308
order parameter 5, 336

p
parallel transport 559
particle horizon 394, 399
path integral formalism 346, 348
Peccei-Quinn symmetry, *see* axion PQ 356
peculiar velocity 386
penetration depth 337
perfect fluid 384, 427, 429, 430, 446
Peskin-Takeuchi S, T parameter 311
phantom energy 533
phase field 32
phase rotation 126
phase transition 17, 18, 418
– first order 404, 418, 420
– second order 418
photo-disintegration 411
photon decoupling 395, 417, 435
– time 416
photon temperature 409
pion decay constant 29, 360
Planck energy 267
Planck satellite *see* CMB Planck 519
Planck's formula 389, 395, 455
plasma frequency 373
plasmon 73, 145
PMNS matrix 73, 116, 120
pNGB: pseudo-Nambu-Goldstone Boson 29
Poincaré group 194
Poisson equation 429, 430, 484
polytrope gas 545
Pontecorvo-Maki-Nakagawa-Sakata matrix, *see* PMNS matrix 73
Pontryagin charge 338, 557
power spectrum 438, 442, 447, 448, 462
– CMB, *see also* CMB anisotropy 454
– large scale structure 437
Primakov effect 494
proper distance 380, 396, 399
proton decay 171, 172, 175, 185, 187, 285, 286
pseudo Nambu-Goldstone Bosons 24

q
quadrupole
– component 470
– matrix 487
– moment 485, 487
quantum anomaly, *see* chiral anomaly 158
quantum tunneling 350
quark confinement phase 395
quark mass
– scale dependence 173
quark number, *see* baryon number 158
quartic coupling
– *see* Higgs 32
quasar 417, 466

quintessence
– attractor 530
– energy density 529
– freezing model 524, 529–531
– model 523
– scaling solution 525
– slow rolling 524
– slow rolling field 531
– thawing model 530, 531
– tracker 523, 528–530
– tracker solution 526
– tracking condition 526, 529
– tracking potential 528

r
R-parity, 229, *see* SUSY models 308
r-process 516
radiation dominated 392, 393, 397, 399, 414
radiative correction 17
radioisotope 516, 518
radion 296, 300
21 cm radiwave 476
razor analysis 258
reactor experiment 111
– Daya Bay 118, 119
– energy spectrum 147
– KamLAND 87, 109, 111, 117, 119
– – data 113
recombination 395, 414, 415, 417, 454, 464, 466, 468
red giant 365, 518
Red Out 533
red shift 385, 386
– of particles 386
reduced Planck mass 292, 421
refraction 101
– index 101
reheating 418–420
reionization 417, 466, 468
relics of big bang 489
renormalization group equation 11, 14, 160, 170, 171, 173, 185, 233–235, 244, 247, 283, 284
resonant density 103
resonant mixing 102, 104
RGB: red giant branch 364, 365, 367
RGE, *see* renormalization group equation 11
ρ: neutral to charged current ratio 9, 20
Ricci scalar curvature 293, 378, 561
Ricci tensor 378, 561
Riemann tensor 561
right-handed gauge boson *see* W_R 69
Robertson-Walker metric 379, 396, 427, 446
rotation curve
– Milky Way 499
– spiral galaxy 476, 477
RSD: Redshift Space Distortion 539
Rydberg-atom 374

s
Sachs-Wolfe effect 458
Saha’s equation 416
Sakharov’s three conditions 402, 404
SB, *see* neutrino super-beam 124
Scherk-Schwartz mechanism, *see* symmetry breaking by boundary condition 277
Schwarzschild radius 400
SD: spin dependent cross section 500
SDSS 426
seesaw mechanism 61, 63, 67–69, 76, 121, 126, 407
shear 479, 486, 487
– field 482, 488
– matrix 482, 484, 487
– measurement 479
– pattern 486
– – B-mode 486
– – E-mode 486
– reduced complex – 488
shift transformation 32
SI: spin independent cross section 500
Silk damping 435, 462
– length 463
Silk mass 463
simplified model 249
simply connected 553
– Lie group 424
Sine-Gordon equation 332
SL(2,C) 194
slow rolling 418–420
slow rolling parameter 525
SM: Standard Model 2
SN1987A 73, 87, 97, 136, 149–151
sneutrino 229
SO(10) 69, 174
solar neutrino 93
– ^7Be 109
– ^8B 92, 107, 108
– CNO cycle 95, 99
– energy spectrum 96, 98
– Ga experiment 98
– GALLEX/GNO 98
– Homestake experiment 94
– LMA:large mixing angle solution 106
– pp-chain 94, 95
– puzzle 79, 94, 100
– SAGE 98
– SMA:small mixing angle solution 106, 109

solar neutrino (*contd.*)
– SNO 99, 100
– SNU 96
– SSM, *see* Standard Solar Model 93
– SuperK 97
– survival probability 107
solar term 122
solitary wave, *see* soliton 329
soliton 329, 551
sound horizon 432, 443, 444, 461, 464
sound velocity 431–433, 436, 460
spallation 91, 113, 119
specific heat 507
sphaleron 349, 350, 404, 406, 407
splitting function 42
SQUID 374
M: sequential Standard ModelI 30, 180, 181, 324
M:standard solar modelI 93, 94, 97–100
standard candle 513–515, 539
stellar
– cooling 73, 152
– cycle 149, 150
– evolution 62, 138, 363
– evolution model 73, 97, 137, 141
Stokes parameter 468, 469
strong CP problem 329, 350, 355–357
structure formation 395, 414, 425
SU(2) 2
SU(5) 160
SUGRA, *see* supergravity 193, 234
SUGRA GUT 231
SUGRA scale 233
super algebra, *see* graded algebra 199
super particle *see* supersymmetry sparticle 225
super Poincaré algebra 201
Super-Kamiokande 88
super-Weyl scale invariance 232
superconductivity 28
supergravity 157, 193
SuperKamiokande *see* neutrino oscillation SuperK 80, 88
supernova, *see also* SN1987A 73
– type Ia 138, 143, 513, 537
– type II 138, 146
superspace 200
supersymmetry 14, 18, 19, 59, 155, 157, 174, 184–186, 189, 225
– action 210
– chiral superfield 203, 208, 209, 211, 214, 218, 220, 223
– – kinetic energy 212
– – left-handed – 204–206
– – products of 206
– – right-handed – 204, 206
– covariant derivative 202, 209, 218, 219
– D-term 219, 233
– – breaking 221, 222
– F-term 216, 233
– – breaking 221, 223, 224, 231
– fermion mass term 213
– gauge interaction 217, 218
– gauge transformation 207–209, 217
– higgsino 187
– Lagrangian 215, 220
– non-Abelian interaction 219
– off-shell representation 220
– on-shell representation 220
– operator 198, 199, 202, 221
– soft breaking 224
– sparticle 185
– spontaneous symmetry breaking 221
– superfield 192, 201, 202
– superfield operator 200
– supermultiplet 192
– superpotential 213–216, 218, 219, 221, 222
– U(1) transformation 217
– vector superfield 206, 218, 220
– – kinetic energy 216, 218
– Wess-Zumino gauge 208, 209, 220
– Yukawa interaction 213
surface of last scattering 395
SUSY, *see* supersymmetry 14
SUSY GUT 173, 185, 187
SUSY models 225
– AMSB 27, 232, 245
– canonical SUSY scenario 249
– chargino 228, 238
– – search 253
– CMSSM 27, 231, 260
– CNMSSM: constrained next MSSM 232, 241
– decoupling regime 27
– gaugino 237, 242
– – mass 243
– GGM 231, 241
– gluino
– – search 255
– GMSB 27, 231, 241, 243, 249
– – features 244
– – mass formula 242
– Goldstino 228, 244, 245
– gravitino 242, 244
– – search 262
– higgsino 187, 228
– long-lived sparticle 252

– LSP 229, 244, 250, 494
– – search 255
– m_0, $m_{1/2}$ 231
– $m_0 - m_{1/2}$ plane 260, 266
– mass spectrum 241, 245, 247–250, 259, 263
– MSSM, *see* MSSM 226
– mSUGRA 231, 234, 235, 242
– – mass formula 237
– μ problem 241
– natural SUSY 263, 265
– neutralino 228, 238
– – search 253
– NLSP: next LSP 244, 252
– nMSSM: next MSSM 241
– photino 228
– pMSSM: phenomenological MSSM 232
– R-hadron 261
– R-parity 229, 261, 308
– searches 248, 263, 264
– selectron search 252
– sfermion mass 243
– singlino 241
– slepton 239
– – search 250
– smuon search 251, 252
– sneutrino 229
– soft breaking 226, 231, 235, 237, 241, 243, 244, 246
– sparticle 229
– – mass spectrum 235
– split SUSY 28, 232
– squark 239
– – search 255
– stau 244
– – search 250
– stop 25, 26, 240
– – mass 26, 27
– – search 260
– symmetry breaking 231
– – in extra dimension 245
– – scale 243, 244
– wino 238
– zino 228
symmetry
– SU(2) 2
– SU(2)×U(1) 4
symmetry breaking
– by boundary condition 277
synchronous gauge 562, 563

t
t'Hooft-Polyakov's monopole 424
tan β 8, 22–25, 55–57
technicolor 29, 184
– condensate 29
– extended – 30
– LSTC, *see* Low Scale Techni-Color model 30
– techni-fermion 29
– techni-gluon 29
– techni-pion 184
– techni-rho 30
– walking – 30
tensor spherical harmonics 469, 565
thermodynamics 389
θ vacuum 329, 347, 348, 352–356
Thomson scattering 417, 463, 466, 467, 469
time delay of light 479
top condensate model 28, 184
topological charge, *see* topological number 334
topological defect 278, 424
topological number 334, 339, 345
topological structure 334, 337
topology 552
torsion experiment 271, 315
transition magnetic moment 72
translation operator 200
transverse polarization 7
tri-bimaximal 120
triangle anomaly 158, 159, 228, 352
triangle anomaly *see also* chiral anomaly 159
triviality 12, 14
twisted ribbon 334
two component formalism 193
two-point function 456

u
U(1) 337
U(1) problem 355, 356
ultraviolet cutoff 283
underground detector 87
unitarity 12, 13, 20, 23, 67, 115, 116
unitary condition 122
unitary triangle 116
universal extra dimension *see* extra dimension UED 308
unnatural 18, 68, 183
UV completion 33, 38

v
V+A interaction 69, 128
V-A interaction 69, 128
vacuum expectation value 5, 8, 11, 16, 21, 22, 29, 34, 53
vacuum stability 10
vector boson fusion 41, 53

VEV, *see* vacuum expectation value 5
virial mass 477, 478
virial temperature 544
virial theorem 477, 543
visible sector 231, 232, 236, 241, 242, 244, 246
– scale 243
vortex 334

w
W boson fusion 39, 40, 45, 46
W' boson 175, 179–181
W' boson *see also* W_R 175
Ward-Takahashi identity 353
weakly interacting massive particle, *see* WIMP 493, 494
Weinberg angle 4, 9, 359
– in GUTs 157, 167, 185
Weyl
– field 193, 206, 213
– representation 193
– spinor 62, 193, 194, 196, 197, 200, 205
white dwarf 364, 365
Wilson phase, *see* Aharonov-Bohm phase 290
WIMP 493, 494, 496, 497
WIMPs wind
– annual modulation 504
– cross section 500
– daily modulation 505
– detection 498, 505
– detector
– – CDMS 507, 508
– – Xenon 508
– – XENON100 510, 511
– mass density 498
– recoil energy 501
winding number 329, 337–339, 341, 342, 344–346, 348, 357, 554
wine bottle potential 331, 359
WKB formalism 346
would-be Goldstone bosons 24
would-be-Goldstone-boson 24, 39
W_R boson 69, 174
– search 179
luminosityI
45

x
X-ray emitting cluster 478
Xenon
– electro luminescence 510
– scintillator 509
– two phase operation 509, 510

y
y-parity *see* extra dimension Z_2 parity 275
Yukawa coupling 6, 17, 26, 30, 35, 38
– constant 11
Yukawa interaction 4, 37

z
Z' boson 175–178
– mass 178
– search 178
zenith angle distributions 80, 100
zero-point energy 191
Zwicky 475